IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 4, 2020, accepted August 21, 2020, date of publication September 2, 2020, date of current version September 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3021211

On Gaze Deployment to Audio-Visual Cues of Social Interactions

GIUSEPPE BOCCIGNONE, VITTORIO CUCULO, ALESSANDRO D'AMELIO, GIULIANO GROSSI, AND RAFFAELLA LANZAROTTI

PHuSe Laboratory, Dipartimento di Informatica, Universitá degli Studi di Milano, 20133 Milan, Italy

Corresponding author: Alessandro D'Amelio (alessandro.damelio@unimi.it)

This work was supported in part by the University of Milano under Grant PSR 2019.

ABSTRACT Attention supports our urge to forage on social cues. Under certain circumstances, we spend the majority of time scrutinising people, markedly their eyes and faces, and spotting persons that are talking. To account for such behaviour, this article develops a computational model for the deployment of gaze within a multimodal landscape, namely a conversational scene. Gaze dynamics is derived in a principled way by reformulating attention deployment as a stochastic foraging problem. Model simulation experiments on a publicly available dataset of eye-tracked subjects are presented. Results show that the simulated scan paths exhibit similar trends of eye movements of human observers watching and listening to conversational clips in a free-viewing condition.

INDEX TERMS Audio-visual attention, gaze models, social interaction, multimodal perception.

I. INTRODUCTION

Consider a clip displaying social interactions, in particular a conversational clip (audio and video): the chief concern of this article is to model the deployment of attention through gaze by a human subject who is viewing and listening to the clip.

Why should this research problem be relevant beyond its merits?

One straightforward reason lies in the classic data mining hurdle. YouTube, Twitch, Facebook Live contain myriads of such clips [1], [2]. Also, large-scale multimodal data conveying social interactions from non-laboratory settings are being increasingly employed to analyse behaviours, emotions, and interactions in real-life situations [3]. It goes without saying, the processing of large spatio-temporal data from multiple media in different contexts is a mind-blowing engineering challenge: spotting sharable highlights, capturing socially relevant events, generate value-based summaries to facilitate browsing and skimming. All such problems call for an ability that is germane to the successful performance of any cognitive task: the ability to predict and to select where the most meaningful and task-relevant information is to be found in the sensory input.

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

A less evident, albeit earnest need takes root in the challenge of ''subject's mining'': the computational inference of subject's traits, or expertise, or even expectations from attentive behaviour. Much can be gained indeed by analysing the ''mind's eye'' conduct of a subject who scrutinises and forages on the behaviour of other subjects involved in social interactions [4]–[7].

In a nutshell, the research problem addressed here is relevant beyond its peculiar interest because it complies with a quest for parsimony. Under a variety of circumstances, what *prima facie* might come across as a conundrum of diverse engineering problems, boils down to the modelling of one and only skill: the effective deployment of attention that organisms have evolved to promote survival and well-being. Surprisingly, the dynamics of deployment has been hitherto overlooked in computational approaches.

A. PROBLEMS AND CHALLENGES

Throughout our lives, we are bond to unfalteringly sample the environment. Moment-by-moment we strive to answer the question: *Where to look next?* Attention guides our gaze to the appropriate location of the scene and holds it in that location for the deserved amount of time given current processing demands [8].

In doing so, like other animals with as diverse evolutionary backgrounds, we exhibit a consistent pattern of

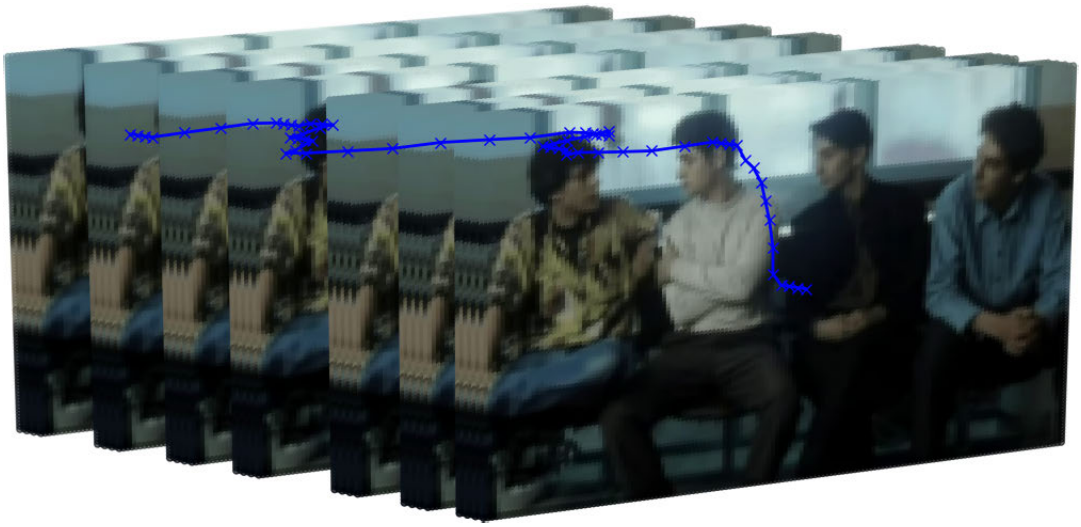


FIGURE 1. Gaze deployment recorded from a human subject who is viewing and listening to a conversational clip. Gaze position in time is rendered by overlapping the raw data recorded along an eye-tracking session on a representative excerpt of video frames. The trajectory unfolding in time is characterised by area-concentrated phases that alternate with large distance relocations between regions attracting attention.

eye movements. To illustrate at the finest ''resolution scale'' the signature of gaze dynamics, Fig. 1 plots the raw data recording of one subject's gaze. The trajectory of gaze is shown as unfolding in time on an excerpt of subsequent frames: large relocations are followed by local clustering of gaze points.

This pattern has been referred to as a ''saccade and fixate'' strategy [9]. Saccades are the fast movements that redirect the eye to a new part of the surroundings, and fixational movements occur within intervals between saccades, in which gaze is held almost stationary. In dynamic scenes, or ones including observer's movement, fixations are either replaced by or augmented with the smooth pursuit eye movement to keep on the fovea (the central part of the retina) the objects of interest that are moving.

The given tasks or goals determine by and large such pattern [8]. Yet and cogent for the work described here, the pattern is not the unconcerned result of a disembodied process. Nor are the given task and the stimuli properties the only contraints to the perceiver. Subject's gut and feelings matter too: in our daily life we keenly move our gaze to gauge and collect visual information that includes social information, such as others' emotions and intentions [7], [10].

The implicational converse of this state of affairs is that the dynamic pattern springing from this lifelong sampling endeavour provides information about plans, goals, interests, and probable sources of rewards; even expectations about future events [8], [11], personality and social traits [7].

In this perspective, conversational videos have the ecological virtue of displaying real people embedded in a dynamic situation while being relatively controlled stimuli. In the conversational setting, Foulsham *et al.* [12] have shown that observers spend the majority of time looking at the people in the videos, markedly at their eyes and faces, and that gaze fixations are temporally coupled to the person who was talking at any one time. This is not surprising. Visually-mediated social interactions are not exclusive to humans, and have played a significant role very early in the primate lineage: selective pressure is likely to have promoted convergent evolution of social gazing abilities for social group-living animals [10]. Modelling attention in such case entails taking into account the value of social cues. This, in turn, raises the question of whether it be feasible to mine from behavioural data the implicit value of multimodal cues that drives observer's motivation.

Even prior to such urgent quest, the audio-visual nature of these stimuli brings forward the challenge of how gaze is to be guided in the context of multimodal perception (audio and visual). As discussed in Section II, limited work has been devoted to eye guidance in a multimodal setting.

B. OUR APPROACH

The key intuition can be easily grasped at a glance by going back to Fig. 1. The trajectory of gaze unfolding in time can be best described, at the phenomenological level, as one kind of biased random walk that takes place at different scales: the fine scale of area-concentrated phases within valuable ''information patches'' (*exploitation*), that alternates with the coarse scale of large distance relocations between patches (*exploration*), whatever the precise rules that control them. Thus, the portrait of Fig. 1 boils down our chief research problem to two crucial questions: *What* defines valuable a patch? *How* is gaze guided within and between patches?

The answer is formalised in a model for eye guidance that relies on a simple idea: we consider gaze trajectories as traced by a composite forager, chasing up resources that are patchily distributed (cfr. Fig. 1). Foraging is a general term that includes where animals search for food and which

sorts of food they eat [13]–[15]. A composite forager is one capable of switching the scale of the foraging walk from within-patch exploitation to large between-patch relocations or vice versa [15]. In our case, the forager is a stochastic one, and either regime - exploitation or exploration - is accomplished via a biased Brownian walk, precisely an Ornstein-Uhlenbeck (OU) process, tuned at the appropriate scale. The bias is provided by the audio-visual patches that moment-by-moment appear relevant (rewarding) within the multimodal landscape. The idea of exploiting the foraging framework has gained currency in the attention literature (cfr. Table 1), reckoned more than an informing metaphor [16].

TABLE 1. Relationship between multimodal attention and foraging.

Audio-visual attentive processing	Patchy landscape foraging
Perceiver	Forager
Perceiver's gaze shift	Forager's relocation
Audio-visual object/event	Patch
Audio-visual object/event selection	Patch choice
Deploying attention to object/event	Patch handling
Disengaging from object/event	Patch leave or giving-up

Technically, as depicted in Fig. 2, model input is represented by the audio-visual stream together with eye-tracking data. We exploit the publicly available dataset presented in [17], collecting data of eye-tracked subjects attending to conversational clips.

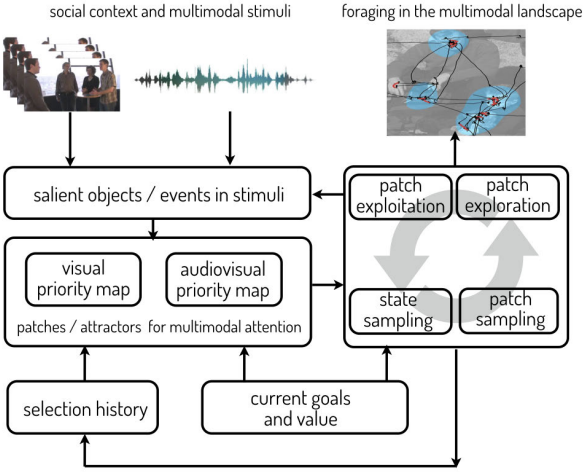


FIGURE 2. Gaze deployment as foraging in a multimodal landscape. Model input is represented by multimodal stimuli that convey social content; the output is represented by a composite (local/global) foraging walk. Value-based patches are sampled from priority maps and integrate different sources of selection bias in a socially valuable context. The audio-visual scene social content drives perceiver's (internal) value that, in turn, guides the sampling of relevant patches. The perceiver's gaze continuously switches between local patch exploitation and between-patch global relocation. Gaze dynamics is that of a spatial Ornstein-Uhlenbeck process, which is performed at two different scales, local and global.

At the pre-attentive stage, inference is performed to obtain dynamic value-driven priority maps resulting from the competition of visual and audio-visual events occurring in the scene. Their dynamics integrates the observer's current selection goals, selection history, and the physical salience of the items competing for attention. The free-viewing task given to subjects allows for dynamically inferring the history of their ''internal'' selection goals as captured by the resulting attentive gaze behaviour. From priority maps a number of attractors are sampled in the form of value-based patches suitable to bias the forager's walk. The attentive stage involves trading between local patch exploitation and landscape exploration through relocations across patches. This is achieved by switching the OU process at different scales. The trading rules stem from stochastic approaches to optimal foraging theory [14].

C. MAIN CONTRIBUTIONS

The novel contributions of this article lie in the following.

1) The proposed model addresses the active sensing of multimodal stimuli (audio and visual). Surprisingly enough and to the best of our knowledge, there is not much tradition in the computational modelling of this problem.
2) Attention deployment is reformulated as a stochastic foraging problem. Albeit unconventional, this choice allows a parsimonious approach to cope with both the *what* and *how* problems that ground active sensing, the *how* problem being hitherto neglected.
3) Gaze dynamics succintly relies upon one and only OU stochastic process that is apt to switch between different scales of diffusion. This solution accounts for the variability problem of the perceivers in a simpler way than some attempts based on more cumbersome mathematical tools (e.g., Lévy flights). A side consequence is to allow a concrete step towards the unified modelling of different kinds of gaze shifts, a recent trend in eye movement research.
4) The foraging framework is exploited for a seamless but principled integration of attentional control mechanisms that are modulated by value and rewards. In particular it is shown how implicit social reward as elicited by multimodal conversational clips can be inferred and exploited in the loop. Value and reward are seldom considered in the computational models of attention.
5) Different from the current propensity towards end-to-end approaches, the model-based behavior of gaze deployment provides an explainable account. This is important if the approach is to be used in a subject's mining context (for example, inferring socially-aware psychological traits of the perceiver or atypical development in the appraisal of social cues).

These results grow out from the efforts spent to overcome some of the limitations of current approaches to model attention. Such limitations are overviewed to some detail in the following Section II. The overview motivates the rationales behind model's architecture, which is presented in Section III. Its formalisation is developed in Sections IV and V. Section VI-A presents simulation results and their analysis. Conclusions are summarised in Section VII.

II. BACKGROUND, RELATED WORK AND LIMITATIONS

We proceed now to set up a minimal formalism needed to outline the necessary background to the work presented here and to compare with the state-of-the-art.

Early studies on gaze behaviour and attention [18], [19] made clear that in this matter three factors are to be taken into account: the task or goal $\mathcal{G}$, the stimuli $\mathcal{S}$, and the perceiver $\mathcal{O}$. Overt attention deployment as instantiated through the unfolding of gaze shifts involves two main processes: i) perception, by which $\mathcal{O}$ processes sensory information and makes inferences to set up a representation $\mathcal{W}$ capturing salient aspects of the world; ii) action $\mathcal{A}$, by which $\mathcal{O}$ chooses how to sample the world to obtain useful sensory information.

The perceptual process can be formalized in terms of an ideal perceiver model which makes task-relevant inferences. The perceiver $\mathcal{O}$ uses the sensory input $\mathcal{S}$ (visual or audio-visual, for example) together with a knowledge of the properties of the task $\mathcal{G}$ and the world, as well as features of the sensors at hand. The process of selecting an action uses both the observer's inferences and knowledge of the goal $\mathcal{G}$ to determine the next movement, i.e. where to orient the eyes. Action execution leads to new sensory input $\mathcal{S}'$. This closes the active sensing loop of perception and action.

In brief, consider time instants $t < t'$, where $t'-t = \delta t$ is an arbitrary time step. Assume that at time t the perceiver's gaze centers the focus of attention (FoA) at location $\mathbf{r}_F(t)$, (subscript F explicitly links location to the FoA). Then, the goal-driven action/perception cycle performed by $\mathcal{O}$ boils down to the iteration of the following steps. Under goal $\mathcal{G}$ and current sensory input $\mathcal{S}(t)$:

Step 1: Infer the current perception of the world $\mathcal{S}(t) \rightarrow \mathcal{W}(t)$ when gazing at $\mathbf{r}_F(t)$;
Step 2: Sample the appropriate motor action/decision $\mathcal{A}(t)$ depending on $\mathcal{W}(t)$;
Step 3: Sample the gaze shift $\mathbf{r}_F(t) \rightarrow \mathbf{r}_F(t')$, depending on $\mathcal{A}(t)$, $\mathcal{W}(t)$.

In a nutshell, the eye guidance loop answers the very question: *Where to look next?* The "where" part (Step 1) concerns the selection of *what* to gaze at - features, objects, actions - and their location within the scene; the "next" part (Steps 2 and 3) involves *how* we gaze at what we have chosen to gaze. The latter crucially brings in the unfolding dynamics of gaze deployment.

The gist of the discussion that follows lies in that, by and large [20]–[23], the computational modelling of visual attention has hitherto been concerned with Step 1 (*what*): deriving a representation $\mathcal{W}$. As a matter of fact, it is surmised that the perceptual representation $\mathcal{W}$ is *per se* predictive of human fixations. Steps 2 and 3 (*how*) are seldom taken into account.

A sober scrutiny of the literature shows that this attitude instantiates in a number of hindrances occurring in all steps and concerning the main factors $\mathcal{G}$, $\mathcal{O}$, $\mathcal{S}$.

1) PROBLEMS WITH $\mathcal{W}$: LEVELS OF REPRESENTATION

Consider the mapping $\mathcal{S} \rightarrow \mathcal{W}$ (Step 1). The guidance of gaze deployment is likely to be influenced by a hierarchy of representational levels. Plausible ones to account for are [24]: 1) *salience*, 2) *objects*, 3) *values*, and 4) *plans*.

Up to this date, as stigmatised in many studies [20]–[23], [25], [26], the majority of computational models have epitomized $\mathcal{W}$, the perceptual representation of the world, in the form of a spatial saliency map, which is mostly derived bottom-up (early salience) on the way paved by Itti *et al.* [27].

The weakness of the bottom-up approach has been largely weighed up in the visual attention realm [20], [24], [26]. To overcome this pitfall, early saliency can be top-down tuned to improve fixation prediction when dealing with objects [28], faces [29], text regions [29], [30] or contextual cues, e.g., the scene gist [31]. Indeed, the success of deep networks exploiting convolutional filters that have been learned on other tasks, for instance object recognition, provides practical evidence of the usefulness of high-level image features for prediction purposes [32], [33]. Despite of this heuristic addition of high-level processing capabilities, these are still referred to as saliency models with some lack of clarity [21]–[23], [32], [34].

2) HOW TO DEFINE $\mathcal{G}$: THE MANY FACETS OF GOALS

As a matter of fact, in the real world gaze is not generically deployed to objects but allocated to task-relevant objects [24], [25], [35]. The recent theoretical perspectives on active/attentive sensing [36] promote the idea that the ultimate objective of the active sensing loop (Steps 1-3) should be to maximise via exploration the long term total rewards and to gain additional knowledge about the environment. Cogently, this endeavour recalls that of animals foraging for food. Animals are likely to choose actions that not only take them closer to known food sources but also yield information about potential new sources [36], [37].

Yet, defining what is a goal is far from evident. The dichotomy between top-down and bottom-up control assumes the former as being determined by the current "endogenous" goals of the observer and the latter as being constrained by the physical, "exogenous" characteristics of the stimuli (e.g.,flashes of light, loud noises, independent of the internal state of the observer). The construct of "endogenous" attentional control is subtle since it conflates control signals that are "external", "internal" and selection history (either learned or evolutionary inherited), which can prioritise items previously attended in a given context. To discuss thoroughly this point would carry us deep into the study of the complex interaction between cognition and emotion [38]. A few words must here suffice.

If the ultimate objective of the attentive perceiver is total reward maximisation, one is urged to distinguish between "external" rewards (incentive motivation, e.g, monetary reward) and reward related to "internal" value. Most important for the work presented here, the latter has different

psychological facets [39] including affect (implicit ''liking'' and conscious pleasure) and motivation (implicit incentive salience, ''wanting''). Indeed, the selection of socially relevant stimuli by attention has important implications for the survival and wellbeing of an organism, and attentional priority reflects the overall value and the history of such selection [40]. Indeed, the crude top-down vs. bottom-up taxonomy of attentional control should be adopted with the uttermost caution (cfr., [20], [41]).

3) THE NEGLECTED PERCEIVER: BIASES, VARIABILITY, IDIOSYNCRASY

To date, the vast majority of models have largely ignored the perceiver. Still, when considering the *how* component (Steps 2 and 3), the ''$\mathcal{O}$ factor'' is cogently brought in.

On the one hand, regardless of the perceptual input, scan paths exhibit both systematic tendencies and notable inter- and intra-subject variability. Systematic tendencies or ''biases'' in oculomotor behaviour can be thought of as regularities that are common across all instances of, and manipulations to, behavioural tasks [42], [43]. One remarkable example is the amplitude distribution of saccades and microsaccades that typically exhibit a positively skewed, long-tailed shape [20], [42]–[44].

As to variability, when looking at natural images, movies [44], or even dynamic virtual reality scenes [45] under a free-viewing or a general-purpose task, there is a small probability that two observers will fixate exactly the same location at exactly the same time. Such variations in individual scan paths (as regards chosen fixations, spatial scanning order, and fixation duration) still hold when the scene contains semantically rich ''objects'' and can become idiosyncratic [20].

Recent studies examined the variability of eye movements between observers distinguishing which characteristics are stable and reliable, and therefore should be treated as a trait of the observer rather than ''noise'' [46], [47]. Guy *et al.* [7] have shown that the amount of time subjects fixate on others' faces (face-preference) varies between individuals in a reliable manner. Biases and variability have been considered a nuisance rather than an opportunity. Nevertheless, beside theoretical relevance for modelling human behavior, the randomness of the process can be an advantage in computer vision and learning tasks [48].

There are few notable exceptions to this current state of affairs, [49], [50], [51], [52]. Variability and bias have been explicitly addressed from first principles in the theoretical context of Lévy flights [53], [54]. Interestingly enough, this direction too leads to treating visual exploration strategies in terms of *foraging* strategies [16], [30], [55]–[57]. In certain circumstances, uncertainty may promote almost ''blind'' visual exploration strategies [43], [58], much like the behaviour of a foraging animal exploring the environment under incomplete information [14].

4) DEFINING $\mathcal{S}$: THE MULTI-SENSORY CHALLENGE

Humans are multi-sensory perceivers. We are capable of attentional behaviour on multimodal stimuli, for example those mixing visual and audio stimuli, $\mathcal{S} = \{\mathbf{I}, \mathbf{A}\}$, where $\mathbf{I}$ is a frame sequence and $\mathbf{A}$ an audio signal. Whilst attentional mechanisms have been extensively explored for vision systems, there is not much tradition as regards models of attention in the context of sound systems [59].

Mutual influence between speech and visual perception, markedly, face perception, is a long debated and well known issue. The link between perceiving speech and perceiving faces has been demonstrated in both behavioural and physiological experiments, e.g., [60]–[64]. The McGurk effect [60] is one celebrated example of audio-visual speech perception, where visual inputs can even override the veridical inputs of the auditory system. Another example is the way people routinely use information provided by the speaker's lip movements to help understand speech in a noisy environment [61], [62]. Watching the lips move in silent video clips activates areas in the auditory cortex that are activated when people are perceiving speech [63]; conversely, when listeners pay attention to a voice that they associate with a specific person [64], this activates areas not only for perceiving speech but also for perceiving faces (face fusiform area, FFA). Van der Burg [65] provided evidence that audio-visual synchrony guides attention in an exogenous manner in adults. However, it remains unclear how multimodal scenes are represented in the brain [66] and there is no comprehensive framework to explain our abilities in multimodal attention.

As to computational models, much like visual attention, the dichotomy between top-down and bottom-up control has been assumed in the auditory attention field of inquiry. Since the seminal work by Kayser *et al.* [67], efforts have been spent to model stimulus-driven attention by computing a visual saliency map of the spectrogram of an auditory stimulus (see [59] for a comprehensive review). In this perspective, the combination of both visual and auditory saliencies supporting a multimodal saliency map that grounds multimodal attention becomes a viable route [68], [69].

Seminal work on multimodal saliency has been done by Coutrot and Guyader [70]–[72], where static and dynamic low-level visual features were combined with semi-automatically segmented object-based cues (such as faces and annotation of body parts). For the audio track of video frames a speaker diarization technique was proposed based on voice activity detection, audio speaker clustering, and motion detection. This information was then combined with visual information to obtain a saliency map. Limitations of that work has been addressed in [73] by providing a framework, which is exploited here to implement the pre-attentive stage of our model (cfr. Section IV and Appendix A). A recent work by Tavakoli *et al.* [74] directly learn the end-to-end mapping for the multi-modal saliency prediction by using a deep neural network instead of relying on a sampling scheme and multiple feature maps.

III. GAZE DEPLOYMENT: OVERVIEW OF THE BASIC MODEL ARCHITECTURE

By taking stock of the limitations highlighted in the discussion above, we next lay down the proposed model of gaze deployment.

The general problem may be stated as follows. The dynamic multimodal landscape $\mathcal{W}(t)$, namely the world as perceived by subject $\mathcal{O}$, is a ''patchy'' environment. Patches are clumps of audio-visual information to which gaze is deployed. The perceiver scrutinises ''items'' within a patch and, at any time t, makes action decisions $\mathcal{A}(t)$ as to: 1) which patches are to be spotted; 2) when to leave the patch currently visited for focussing on a new patch. In this endeavour, the unfolding of gaze deployment, $\mathbf{r}_F(t) \rightarrow \mathbf{r}_F(t')$, alternates between scanning the patch, for probing and exploiting the chunks of information locally available, and longer, explorative relocations between patches.

To frame such problem we make a number of assumptions.

A1 The unfolding of gaze deployment in time is best described as a stochastic process, namely a biased random walk of a forager over the changing landscape (cfr. Fig. 1).

The landscape $\mathcal{W}(t)$ generated by $\mathcal{O}$ from the audio-visual stream $\mathcal{S}(t) = \{\mathbf{I}(t), \mathbf{A}(t)\}$ is inherently stochastic and the observer has partial information, since patches may change unpredictably in time. Further, as discussed in Section II, we need to take into account $\mathcal{O}$'s variability and biases. Interestingly enough, the reformulation of attention in terms of foraging theory goes beyond the informing metaphor. There is substantive evidence that what was once foraging for tangible resources in a physical space became, over evolutionary time, foraging in cognitive space for information related to those resources [75]. Such adaptations play a fundamental role in goal-directed deployment of visual attention [16].

A2 The gaze walk can be accounted for by one and only model of oculomotor behavior, namely an Ornstein-Uhlenbeck process; the process acts at different scales, from landscape exploration to local patch exploitation.

Indeed, recent work has been challenging the view that exploration and fixation are dichotomous. Current literature suggests instead that visual fixation is functionally equivalent to visual exploration on a spatially focused scale [76]. In brief, they are two extremes of a functional continuum. Recent experiments confirmed scale invariance in the temporal structure of the larger shifts in gaze position (saccades), which has also been observed in fixational eye movements while the eye is gauging a localized region in the visual field [77].

A3 In a multimodal landscape conveying social content, the forager's random walk for exploration/exploitation is modulated by the value $\mathbf{v}$, which is internally (self-)assigned by $\mathcal{O}$ to socially rewarding items. Value dynamics can be inferred from $\mathcal{O}$'s oculomotor behaviour.

Here, no ''external'' task is assigned to the perceiver; thus, value $\mathbf{v}$ is modulated by the ''internal'' drive towards spotting socially relevant objects/events. In a landscape featuring social content, the most prominent visual objects are likely to be faces and audio objects as represented by speakers' voices [12]. These, eventually, will maximally contribute to the relevant patches within $\mathcal{W}(t)$ that will bias the random walk of the perceiver's gaze.

Under such circumstances, gaze deployment is obtained as follows. Along a *pre-attentive stage*, audio-visual features are derived to assess the likelihood of the spatio-temporal occurence of such events. This provides the basis for setting up time-varying priority maps $\mathbf{L}_\ell$ ($\ell = 1, \cdots, N_\ell$) and for gauging their moment-to-moment value v_ℓ in the context of the scene. From priority maps, a number of value-based patches $\mathcal{P}_p^{(\ell)}$ ($p = 1, \cdots, N_P^{(\ell)}$) are generated.

The *attentive stage* is distilled in the evolution of the gaze state represented by point $\mathbf{r}_F(t) = (x_F(t), y_F(t))^T$ in a continuous 2-dimensional space, at any time $t \geqslant 0$, which sets the focus of attention (FoA). As such, gaze dynamics $\mathbf{r}_F(t) \rightarrow \mathbf{r}_F(t')$ defines a trajectory, which is the realisation $R_F(t) = \mathbf{r}_F(t)$, of a continuous-time stochastic process $\{R_F(t) : t \geqslant 0\}$. From now on, for sake of simplicity and with some abuse of notation, we shall use $\mathbf{r}_F(\cdot)$ for denoting both the process/random variable and its realisation; the same holds for other random variables, unless otherwise specified.

The process is conceived as an OU process operating at two different scales. These parametrise local and global biased random walks so that area-concentrated phases within patches (exploitation) alternate with large distance relocation phases between patches (exploration).

The switch between exploitation and exploration, is provided by a foraging decision resulting from comparing the expected reward gained within currently exploited patch against the average reward that could be gained moving to other patches available within the landscape. If exploration is undertaken, then the choice of a new patch must be made. State switching and patch choice are the behavioural decisions $\mathcal{A}(t)$ available to the forager at time t.

An ancillary assumption of the model presented here relates to the patch exploitation mechanism. In stochastic foraging theory, the time spent within a patch depends on the potential value of a patch. The latter is based on the expected rate, the forager's current expectations on the number of items in the patch and how easy they should be to find, [78]–[80]. In the case of internal goals, it is difficult to exactly define what is an item. For example, consider a patch embedding a speaker's face. Items could either be main facial shape features (eyes, nose, etc.), or action units of facial expressions, or joint lip movements and spoken words, etc. Even if we could count the items, we would not know how many items are processed when gaze is deployed at point $\mathbf{r}(t)$ in the course of local patch exploration; multiple items might be processed in parallel [81].

On this basis, in the same vein of the foraging literature and its applications in perception [16], [55], [81], our model abstracts from the actual mechanisms of specific gaze behaviour within a region of interest under a given task, but isolates some very relevant phenomenological aspects akin to be shaped in statistical terms. This suits our needs, our concern here being with the general view rather than with the details.

Patches and items within the patch, whatever they may represent, are encountered according to a Poisson process. Together with the associated exponential waiting times, they play an important role to relate points of gaze and global/local scene characteristics [82], [83]. Patches are modelled as independent Poisson process generators. Number of items are sampled from a Poisson distribution, which allows to derive a simple law for estimating the instantaneous information gain of the perceiver within the patch and to compare the latter with the average gain achievable over the landscape. This provides a sound basis for deciding when to relocate to another patch and how to choose the next patch to be exploited, namely the actions $\mathcal{A}(t)$ moment-by-moment available to the perceiver.

The control algorithm for gaze deployment is summarised in the *GazeDeploy* procedure outlined in Algorithm 1. Its steps are detailed in the following sections and a Python simulation of the procedure is freely available on GitHub.[1] Figure 3 provides a useful insight of the overall behaviour of the procedure. Given an input conversational clip, summarised as an excerpt of four subsequent frames (top to bottom, left column), the *GazeDeploy* procedure outputs a continuous gaze trajectory as generated by one artificial observer (second column), whilst the third column shows the focus of attention (FoA) set at the corresponding time. In the top, second and bottom rows the simulated observer scrutinises the current speaker, as espected, whilst in the third and fourth raws, a brief glance is deployed either to the woman listening on the left of the scene and to the onset of the hand gesture of the forthcoming speaker. It is worth remarking that one such trajectory is likely to stochastically deviate, to some extent, from those of other observers, either real or artificial. This variability can be appreciated from the fourth and the last columns. They present the time-varying fixation maps computed from a paired number of either artificial observers and actual human observers. Note that when the conversational scene becomes more complex (typically due to people arguing, gesturing, turn-taking, etc.), the maps are characterised by higher spatio-temporal dispersion, a signature of the attention variability between observers. Such uncertainty is captured by both the artificial and actual maps. In such circumstances, indeed, the inter-observer variability grows, and individual observers are likely to be driven by their own expectation and other idiosyncratic factors.

[1]Python code for simulations available at https://github.com/phuselab/GazeDeploy

Algorithm 1 Gaze Control in a Multimodal Landscape

Input: Visual stream $\{\mathbf{I}\}$, audio stream $\{\mathbf{A}\}$, goals $\mathcal{G}$ (internal or external), the duration T to be simulated, the video frame rate FR, the random walk sampling rate fs.
Output: Prediction of gaze
procedure GazeDeploy
$\delta t = \frac{1}{FR}, \quad \delta u = \frac{1}{fs}$
Initialisation of first gaze location $\mathbf{r}(t_1)$ on patch $p(t_1) = j$, with behavioural state $s(t_1) = 0$ {*Exploitation mode*}
for $n = 2$ **to** $\frac{T}{\delta t}$
  {**Preattentive feedforward stage**}
  Compute the current state of the perceptual landscape, in terms of audio-visual priority maps $\{\mathbf{L}_\ell\}$ and distributions $\{\mathcal{L}_\ell(t_n)\}$ (Eqs. 1,2, 3,4)
  {**Value inference**}
  Infer value dynamics $\{v_\ell(t_n)\}$ given all available information $\mathcal{I}(t_1 : t_n)$ up to time t_n, (Eq. 5)
  {**Landscape evaluation**}
  Compute audio-visual patches $\{\mathcal{P}_p(t_n)\}$ as potential value-sensitive attractors
  Compute the expected average gain $Q(t_n)$ from all patches in the landscape (Eq. 28)
  {**Attentive stage**}
  if $s(t_n) = 1$
    {**Exploitation: patch handling**}
    Set the parameters $\boldsymbol{\mu}_p^{(s_t)}, \mathbf{B}_p^{(s_t)}, \boldsymbol{\Psi}_p^{(s_t)}$ for OU sampling according to state $s(t_n)$ and current patch indexed by $p(t_n)$
    while within patch
      {**Exploitation: local gaze shifting**}
      for $j = 0$ **to** $\frac{fs}{FR}$
        Sample the OU gaze relocation $\mathbf{r}_F(t_{n-1}+(j\times\delta u)) \rightarrow \mathbf{r}_F(t_{n-1}+(j+1\times\delta u))$
      end for
      {**Behavioural state sampling**}
      Compute the instantaneous expected gain $g_p(t_{W_p})$ for current patch (Eq. 26)
      Compare current patch gain against the expected average gain Q from the environment (Eq. 29)
      Sample the behavioural state $s(t_n)$ at time $t_n = t_{n-1} + \delta t$ (Eq. 20)
    end while
  else
    {**Exploration: patch-choice**}
    Sample next most valuable attractor $p(t_{n-1} + \delta t)$
    Set the parameters $\boldsymbol{\mu}_p^{(s_t)}, \mathbf{B}_p^{(s_t)}, \boldsymbol{\Psi}_p^{(s_t)}$ for OU sampling according to state $s(t_n)$ and attractor $p(t_n)$
    {**Exploration: relocation gaze shifting**}
    for $j = 0$ **to** $\frac{fs}{FR}$
      Sample the OU gaze relocation $\mathbf{r}_F(t_{n-1} + (j \times \delta u)) \rightarrow \mathbf{r}_F(t_{n-1} + (j + 1 \times \delta u))$
    end for
  end if
end for
end procedure

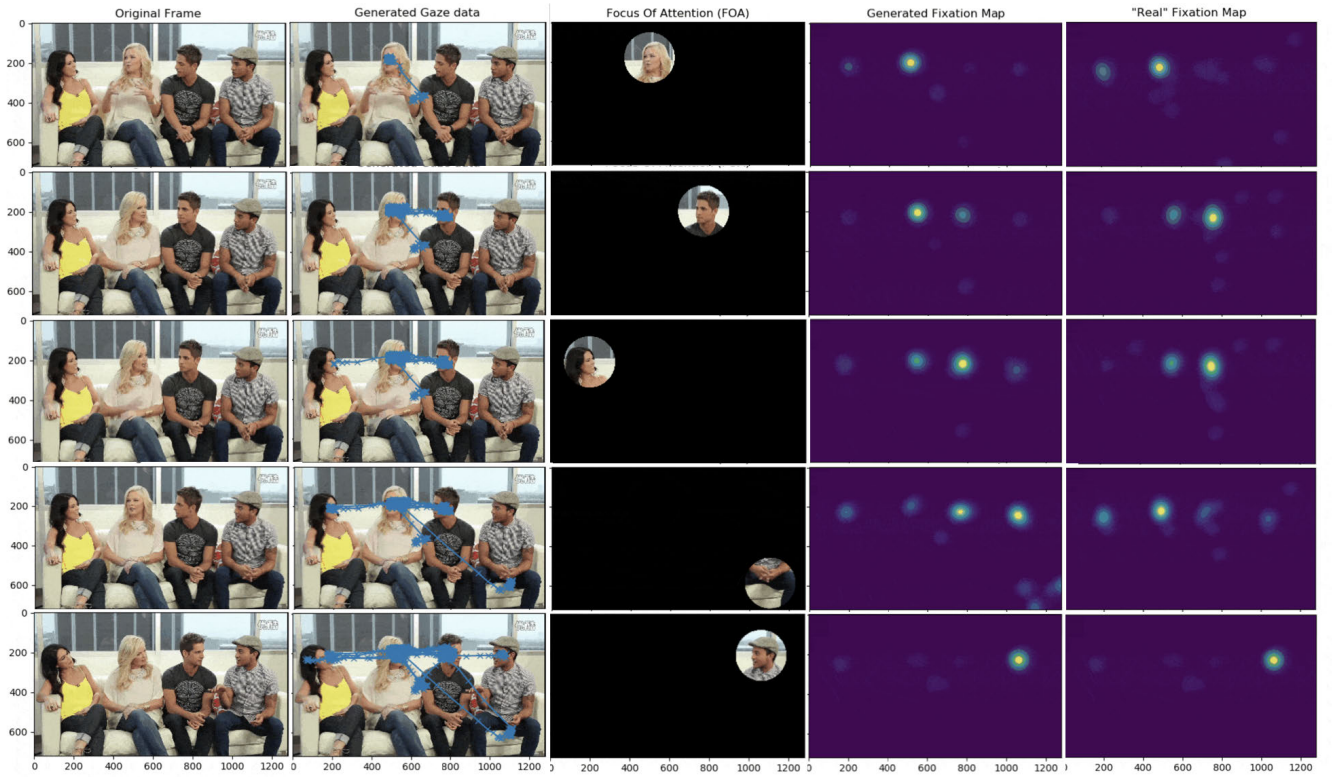


FIGURE 3. The behaviour of the *GazeDeploy* procedure captured through the excerpt of four subsequent frames of a conversational clip. The left-most column summarises the input sequence (top to bottom). The second column displays the output of the procedure, namely the continuous gaze trajectory (graphically overlapped on the input frame) as generated by one artificial observer up to that frame. The third column highlights the focus of attention (FoA) set on the scene. To weigh such individual trajectory in the context of other observers' behaviour, the fourth and right-most columns represent the time-varying fixation maps (a.k.a, heatmaps, attentional maps) computed from a paired number of either artificial observers and actual human observers, respectively.

IV. THE PREATTENTIVE STAGE: PERCEIVING THE AUDIO-VISUAL LANDSCAPE AND ITS VALUE

At the heart of the time-varying, pre-attentive perceptual representation $\mathcal{W}(t)$ lies the concept of priority map. Intuitively, a priority map $\mathbf{L}$ combines top-down (relevance under given goals $\mathcal{G}$) and bottom-up (salience) mechanisms for eye guidance [84]–[87]. More generally, it can be conceived as a dynamic map of the perceptual landscape constructed from a combination of properties of the external stimuli, intrinsic expectations, and contextual knowledge [28], [31]; it can also be designed to act as a form of short term memory to keep track of which potential targets have been attended. As such, the representation entailed by a priority map differs from that provided at a lower level by feature maps $\mathbf{X}$ (or classic salience).

Priority maps are used in our model to sample the audio-visual patches of interest that define the perceiver's landscape. Each patch bears a value inherited from its priority map. Here, rather than shaping value in the form of a map (in a sense, a further instance of a priority map, see [88], [89]), we consider it as a process that moment to moment weighs the relevance of the the different priority maps conditionally on the observer's goal.

A. COMPUTING PRIORITY MAPS

Formally, a priority map $\mathbf{L}$ is the matrix of binary random variables $l(\mathbf{r})$ denoting if location $\mathbf{r}$ is to be considered relevant ($l(\mathbf{r}) = 1$) or not ($l(\mathbf{r}) = 0$), with respect to possible visual or audio-visual "objects" occurring within the scene. Further, $\mathbf{L}(t)$ depends on both current perceptual inferences on feature maps $\mathbf{X}(t)$ at time t and priority $\mathbf{L}(t - \delta t)$ at time $t - \delta t$.

It can be assumed that many such spatially mapped structures contribute to competition, working in parallel across the perceptual field [84]–[89]. To derive the set of priority maps $\{\mathbf{L}_\ell\}_{\ell=1}^{N_\ell}$, N_ℓ being the total number of priority maps, and the related probability distributions $P(\mathbf{L}_\ell)$, the first inferential step concerns the mapping from the multimodal input $\mathcal{S}(t) = \{\mathbf{I}(t), \mathbf{A}(t)\}$ to a set of feature maps $\{\mathbf{X}_\ell\}$. In particular, we are considering the feature maps $\mathbf{X}_\mathbf{I}$ (supporting the low-level saliency map), $\mathbf{X}_{\mathbf{O}_V}$ (visual object-based map), and $\mathbf{X}_{\mathbf{O}_{AV}}$ (audio-visual topographic maps of speaker/non-speakers). Feature maps represent the occurence at a spatial location of the scene of features of interest, namely, generic visual features $\mathbf{F}_\mathbf{I}$, object-dependent visual features $\mathbf{F}_{\mathbf{O}_V}$, and audio (speech) features $\mathbf{F}_{\mathbf{O}_A}$. The computation of feature maps and related distributions relies on previous work [73], which is briefly summarised in Appendix A for the sake of

completeness. Denote for compactness,

$$\mathcal{S}_{VI}(t) = P(\mathbf{X}_{\mathbf{I}}(t) \mid \mathbf{F}_{\mathbf{I}}),$$
$$\mathcal{S}_{VO}(t) = P(\mathbf{X}_{\mathbf{O}_V}(t) \mid \mathbf{F}_{\mathbf{O}_V}),$$
$$\mathcal{S}_{AV}(t) = P(\mathbf{X}_{\mathbf{O}_{AV}}(t) \mid \mathbf{X}_{\mathbf{O}_A}(t), \mathbf{X}_{\mathbf{O}_V}(t), \mathbf{F}_{\mathbf{O}_A}, \mathbf{F}_{\mathbf{O}_V}),$$

the distributions related to the feature maps.

Consider subsequent time instants $t < t'$, where $t'-t = \delta t$ with δt being an arbitrary time step. Define

$$\mathcal{L}_{VI}(t') = P(\mathbf{L}_V(t') \mid \mathbf{L}_V(t), \mathbf{X}_{\mathbf{I}}),$$
$$\mathcal{L}_{VO}(t') = P(\mathbf{L}_V(t') \mid \mathbf{L}_V(t), \mathbf{X}_{\mathbf{O}_V}),$$
$$\mathcal{L}_{AV}(t') = P(\mathbf{L}_{AV}(t') | \mathbf{L}_{AV}(t), \mathbf{X}_{\mathbf{O}_{AV}}),$$

the distributions related to the priority maps. Then, the latter can be estimated as:

$$\mathcal{L}_{VI}(t') = \alpha_V \mathcal{S}_I(t') + (1-\alpha_V)\mathcal{L}_{VI}(t), \quad (1)$$
$$\mathcal{L}_{VO}(t') = \alpha_V \mathcal{S}_{VO}(t') + (1-\alpha_V)\mathcal{L}_{VO}(t), \quad (2)$$
$$\mathcal{L}_{AV}(t') = \alpha_{AV} \mathcal{S}_{AV}(t') + (1-\alpha_{AV})\mathcal{L}_{AV}(t). \quad (3)$$

where α_V and α_{AV} weight the contribution of currently estimated feature maps with respect to previous priority maps, and the $\mathcal{L}_\ell(t')$ are eventually normalised in the $[0, 1]$ interval. In this study, we set $\alpha_V = \alpha_{AV} = 0.8$. This was experimentally determined via ROC analysis with respect to evaluation metrics (cfr. [73]); such value grants higher weight to current information in order to account for changes in the audio-visual stream.

Priority map dynamics requires a prior that can be designed to account for spatial tendencies in the perceptual process. For instance, human eye-tracking studies have shown that gaze fixations in free viewing of dynamic natural scenes are biased toward the center of the scene (''center bias'', [43], [50]), which can be modelled by assuming a Gaussian distribution located on the viewing center $\boldsymbol{\mu}_C$,

$$\mathcal{L}_C = \mathcal{N}(\mathbf{L}; \boldsymbol{\mu}_C, \boldsymbol{\Sigma}_C). \quad (4)$$

B. INFERRING THE VALUE OF PREATTENTIVE INFORMATION

Attentional value is set by the ''internal'' goal (drive) $\mathcal{G}$ towards spotting socially relevant objects/events occurring in the scene. As such, it is a hidden state of the perceiver. The problem we are facing now is to set up an inferential procedure so that, given all available information from the onset of the process up to time t, say $\mathcal{I}(1:t)$, the latent value $\mathbf{v}(t)$ can be estimated,

$$\mathbf{v}(t) \mid \mathcal{I}(1:t) \sim P(\mathcal{I}(1:t)). \quad (5)$$

Information $\mathcal{I}(t)$ should encompass both perceivers' behaviour and stimulus content. Consider that, on the one hand, we know that the actual moment-to-moment deployment of attention over the landscape is the outcome of a value assignment procedure. We assume that the result of attention allocation is summarised through the time-varying heatmap $\mathcal{H}(t)$, which can be easily computed from eye-tracked gaze positions (fixations) of the perceivers [34]. On the other hand, the information available from the stimulus is, at this point, pre-attentively captured via densities $\mathcal{L}_\ell(t)$. Recall that a priority map density $\mathcal{L}_\ell(t)$ can be conceived as a dynamic predictor of potential gaze allocation in space. We surmise that each map contributes to such prediction conditionally on the value it bears for the observer at moment t.

Formally, define $\mathbf{v}(t) = (v_1(t) \cdots v_{N_\ell}(t))^T$ the time-varying random vector of values that are internally assigned to priority map densities $\mathcal{L}_\ell(t)$. Under such circumstances, the mapping $\mathcal{H}(t) = h(\{\mathcal{L}_\ell(t)\}, \mathbf{v}(t))$ can be simply cast in terms of the linear regression equation

$$\mathcal{H}(t) = \sum_\ell v_\ell(t)\mathcal{L}_\ell(t) + \omega(t), \quad (6)$$

which specifies the observers' heatmap $\mathcal{H}(t)$ as the linear combination of predictors (regressors) derived from the stimulus, namely the priority densities $\mathcal{L}_\ell(t)$, perturbed by noise $\omega(t)$. Here, $\mathcal{H}(t)$ is a $2D$ matrix having dimensions equal to the dimensions of the $\mathcal{L}_\ell(t)$ matrices. Eq. 6 specifies a time-varying linear regression, since $v_\ell(t)$ are unknown time-varying coefficients. A straightforward dynamics for the latter is to let $v_\ell(t)$ vary over time according to a random walk, where the value displacement $dv_\ell(t)$ simply amounts to a Brownian displacement $dW_\ell(t)$, i.e. $dv_\ell(t) = dW_\ell(t)$.

Then, the dynamic regression model can be conveniently written in terms of the following vector state-space model:

$$\mathbf{h}(t) = \mathbf{P}(t)\mathbf{v}(t) + \boldsymbol{\omega}(t), \quad \boldsymbol{\omega}(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{R}(t)) \quad (7)$$
$$\mathbf{v}(t) = \mathbf{v}(t-\delta t) + \boldsymbol{\epsilon}(t), \quad \boldsymbol{\epsilon}(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}(t)) \quad (8)$$

where: $\mathbf{h}(t) = vec(\mathcal{H}(t))$ is the observation vector of dimension $|\mathcal{H}| \times 1$, obtained by vectorising matrix $\mathcal{H}$; $\mathbf{P}(t) = \left[vec(\mathcal{L}_1(t)) \mid \cdots \mid vec(\mathcal{L}_{N_l}(t))\right]$ is the matrix whose columns are the vectorised predictors.

The Gaussian disturbances, namely, the process noise $\boldsymbol{\epsilon}(t)$ (with $\mathbf{Q} = cov(\mathbf{v})$) and the observation noise $\boldsymbol{\omega}(t)$ (with $\mathbf{R} = cov(\mathbf{h})$) are both serially independent and also independent of each other.

Online inference of value (Eq. 5) can eventually be performed by solving the filtering problem $P(\mathbf{v}(t) \mid \mathbf{h}(1:t))$ under Markov assumption, where $\mathbf{h}$ is a function of the priority map distributions $\mathcal{L}_\ell$ via the observation/regression in Eq. 7. This way, current goal and selection history effects are both taken into account [41].

C. SAMPLING VALUE-SENSITIVE PATCHES

Priority maps and related values are then used for patch sampling. Patches formalise the concept of multimodal attention attractors and inherit the value from the generating priority maps.

Given a priority map $\mathbf{L}_\ell$, the spatial support of the $N_P^{(\ell)}$ possible patches is computed. Denote $\mathcal{M}_p^{(\ell)} = \{m_p^{(\ell)}(\mathbf{r})\}_{\mathbf{r}\in\mathbf{L}_\ell}$ the map of binary RVs indicating the presence or absence of a patch p. Assume independent patches, within and across priority maps $\mathbf{L}_\ell$. The map of patches generated by $\mathbf{L}_\ell$ is defined

as $\mathcal{M}^{(\ell)} = \bigcup_{p=1}^{N_P^{(\ell)}} \mathcal{M}_p^{(\ell)}$, where $\mathcal{M}_p^{(\ell)} \bigcap \mathcal{M}_k^{(\ell)} = \emptyset, p \neq k$ and the overall patch support map is $\mathcal{M} = \bigcup_{\ell=1}^{N_\ell} \mathcal{M}^{(\ell)}$.

To derive patches from priority maps, we need to estimate their support $\mathcal{M}^{(\ell)}(t) = \{m^{(\ell)}(\mathbf{r}, t)\}_{\mathbf{r}\in\mathbf{L}_\ell}$, such that $m^{(\ell)}(\mathbf{r}, t) = 1$ if $\mathcal{L}_\ell(t) > T_M$, and $m^{(\ell)}(\mathbf{r}, t) = 0$ otherwise. The threshold T_M is adaptively set so as to achieve 90% significance level in deciding whether the given priority values are in the extreme tails of the pdf $\mathcal{L}_\ell$. The procedure is based on the assumption that an informative patch is a relatively rare region and thus located in the tails of the distribution.

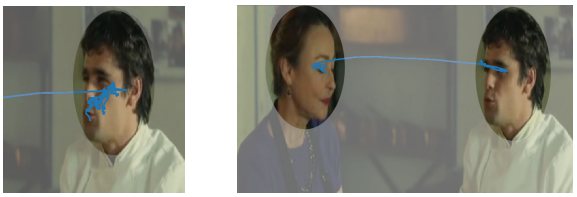
Once the overall support of all patches $\mathcal{M}$ is available, we estimate the parameters defining each patch, namely $\mathcal{P}_p = (\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p, v_p)$ representing its location, shape and value respectively. The value is simply inherited from the generating priority map $v_p = v_\ell$. Location and shape parameters are derived so to provide an elliptical representation of the patch support (patch centre and axes).

V. ATTENTIVE STAGE: THE STOCHASTIC WALK DRIVEN BY THE AUDIO-VISUAL PATCHES

At this point, the input for the attentive stage is available in the form of N_P value-sensitive foraging patches $\mathcal{W}(t) = \{\mathcal{P}_p(t)\}_{p=1}^{N_P}$, with $\mathcal{P}_p(t) = (\boldsymbol{\mu}_p(t), \boldsymbol{\Sigma}_p(t), v_p(t))$, that define the multimodal landscape for the forager's walk.

A. DYNAMICS OF THE WALK

Consider the simple case where a single patch of the viewed scene centered at location $\boldsymbol{\mu}$ (center of mass) serves as an attentional attractor, e.g. the face patch in Fig.4a. The gaze approximately fluctuates (fixational movement) for a time interval around $\boldsymbol{\mu}$.



(a) (b)

FIGURE 4. (a) A face patch serving as attractor of attention, where the gaze deployment in time can be described as a biased 2-D random walk (b) Two face patches representing multiple centers of attraction, with an example of fixation and relocation among patches.

We can idealise the motion of gaze as that of a particle. In Newtonian dynamics the attraction of a particle of position $\mathbf{r}(t)$ pulled towards the location $\boldsymbol{\mu}$ can be described by means of a potential function, a quadratic form $H(\mathbf{r}, t) = \frac{1}{2}(\boldsymbol{\mu} - \mathbf{r}(t))^T \mathbf{B}(\boldsymbol{\mu} - \mathbf{r}(t))$ that controls the particle's direction and velocity $\dot{\mathbf{r}}(t)$; in particular, $\mathbf{B}$ is the 2×2 matrix that constrains the strength of the attraction. In the case that friction is high, particle's velocity is not directly involved and the equation of motion can be written [90], [91]

$$d\mathbf{r}_F(t) = -\nabla H(\mathbf{r}_F, t)dt, \tag{9}$$

with $\nabla = (\partial/\partial x, \partial/\partial y)^T$ the gradient operator applied to the potential and defining the force field $\mathbf{F} = \nabla H$.

When motion is subject to random forces, Eq. 9 generalises to the stochastic differential equation (SDE)

$$d\mathbf{r}_F(t) = \mathbf{B}[\boldsymbol{\mu} - \mathbf{r}_F(t)]dt + \mathbf{D}(\mathbf{r}_F(t))d\mathbf{W}(t), \tag{10}$$

where $\mathbf{B}[\boldsymbol{\mu} - \mathbf{r}_F(t)] = -\nabla H(\mathbf{r}_F, t)$ is the drift term, $\mathbf{D}$ is a 2×2 matrix representing the diffusion parameter. The noise term $\mathbf{W}(t)$ is a 2-D Brownian process that leads to variability around deterministic motion. Simply put, in the stochastic case the particle (gaze) is wandering but being pulled towards the location $\boldsymbol{\mu}$.

Eq. 10 can be easily recognised as a Langevin-type equation. Precisely, the gaze trajectory $\mathbf{r}_F(t), t \geq 0$ is an instance of the 2-D mean-reverting Ornstein-Uhlenbeck (OU) process, where typically $\mathbf{B} = (b_x, b_y)^T$, $\mathbf{D}\mathbf{D}^T = \sigma^2\mathbb{I}$ and $\mathbf{W} = (W_x, W_y)^T$ are independent Brownian processes. Clearly, when $\mathbf{B} = \mathbf{0}$, the drift term is $\mathbf{0}$ and the OU process boils down to the Brownian walk. Eq. 10 can be esplicitly written in the two dimensions as

$$dx_F(t) = b_x[\mu_x - x_F(t)]dt + \sigma dW_x(t), \tag{11}$$
$$dy_F(t) = b_y[\mu_y - y_F(t)]dt + \sigma dW_y(t). \tag{12}$$

Consider the 1-D process on the x coordinate. It is known that for $t \geq 0$, with initial value $x_F(0) = x_0$, the explicit solution of Eq. 11 writes (see e.g. [92], [93]):

$$x_F(t) = x_0 e^{-b_x t} + \mu_x(1 - e^{-b_x t}) + \sigma_x^2 \int_0^t e^{-b_x(t-s)} dW_x(s), \tag{13}$$

and analogously for the $y(t)$ process. The solution can be equivalently written as the conditional sampling

$$x_F(t) \mid x(0) \sim \mathcal{N}(\mu_x + e^{-b_x t}(x_0 - \mu_x), \gamma_x(1 - e^{-2b_x t})), \tag{14}$$

with $\gamma_x = \frac{\sigma^2}{2b_x}$, so that the expected value is $\mathbb{E}[x_F(t)] = \mu_x + e^{-b_x t}(x_0 - \mu_x)$ and the variance ia $var(x_F(t)) = \gamma_x(1 - e^{-2\, b_x t})$. The same holds for the $y_F(t)$ process.

The explicit evolution of x_F in time between 0 and t can be obtained by numerically advancing the particle position with an update equation. This is derived by replacing t in the exact solution (Eq. 13) with $t' = t + \delta t$, δt time units later, and applying the initial condition $x_0 = x_F(t)$:

$$x_F(t') = x_F(t)e^{-b_x \delta t} + \mu_x(1 - e^{-b_x \delta t}) + \sqrt{\gamma_x(1 - e^{-2b_x \delta t})}z(t). \tag{15}$$

In the same way, Eq. 14 writes as the conditional distribution

$$x_F(t') \mid x_F(t) \sim \mathcal{N}(\mu_x + e^{-b_x \delta t}(x(t) - \mu_x), \gamma_x(1 - e^{-2b_x \delta t})). \tag{16}$$

Interestingly enough, Eqs. 15 and 16 can be read as solving Eq. 11 via Monte Carlo simulation, where a sequence of such updates with the realization of the updated position $x(t')$ at the end of each time step is used as the initial position $x(t)$ at the beginning of the next.

Eventually, Eq. 16 and the corresponding one for the $y(t)$ coordinate can be generalised in compact form as

$$\mathbf{r}_F(t') \mid \mathbf{r}_F(t) \sim \mathcal{N}(\boldsymbol{\mu} + e^{-\mathbf{B}\delta t}(\mathbf{r}_F(t) - \boldsymbol{\mu}), \boldsymbol{\Psi}), \quad (17)$$

which represents the general solution to Eq. 10, with $\boldsymbol{\Psi} = \boldsymbol{\Gamma} - e^{-\mathbf{B}\delta t}\boldsymbol{\Gamma}e^{-\mathbf{B}'\delta t}$; $\mathbf{B}$ and $\boldsymbol{\Gamma} = \frac{\sigma^2}{2}\mathbf{B}^{-1}$ are 2×2 matrices and $e^{-\mathbf{M}}$ is the matrix exponential.

Equation 17 describes gaze dynamics towards one point of attraction. In our case, the visual landscape is a time-varying landscape with multiple attractors, the centres of patches $\mathcal{P}_p$. This problem has been partially considered in animal ecology. Breed *et al.* [94] have proposed a multi-state extension of Eq. 17, considering multiple centers of attraction. These centers have unique OU parameters $\boldsymbol{\mu}_i, \mathbf{B}_i, \boldsymbol{\Psi}_i$. However, relocation paths between attractors are not explicitly modelled, which in our case would correspond to the important case of medium/long saccades. Also, along time multimodal patches can vary in number, shape and value.

Harris and Blackwell [95] proposed a flexible class of continuous-time models for animal movement, allowing movement behaviour to depend on location in terms of a discrete set of regions and also on an underlying behavioural state. The diffusion processes that the individual follows while in a particular combination of state and region are, by assumption, OU processes. Thus, for each combination, the parameters of the OU process are specified as, $\boldsymbol{\mu}_i^{(s)}, \mathbf{B}_i^{(s)}, \boldsymbol{\Psi}_i^{(s)}$, for states $s = 1, \ldots, K$, and regions $i = 1, \ldots, L$. The switching process is a continuous- time finite state Markov chain. Its properties are therefore defined by its generator [95], the matrix of instantaneous rates of transition between states observed at short time intervals of length δt. Again, such approach is unfeasible, in our case, where the number of attractors - and, consequently, the number of states- is not known a priori and varies in time.

In our case, we are more truly dealing with two behavioural states that are independent of location: local intensive foraging and extensive exploration. Denote $\{S(t) : t \geqslant 0\}$ a process defined on a binary set $s_t \in \{0, 1\}$ accounting for such behaviour switching process. Its value represents which state of the hidden behaviour is active: foraging, when $s_t = 1$, or exploration when $s_t = 0$ at time t. The regions of attraction are represented by the ensemble of patches $\mathcal{W}(t) = \{\mathcal{P}_p(t)\}_{p=1}^{N_P}$.

In this setting the parameters $\boldsymbol{\mu}_p^{(s_t)}, \mathbf{B}_p^{(s_t)}, \boldsymbol{\Psi}_p^{(s_t)}$ of the OU process are related to a chosen patch p identified through its center location parameter $\boldsymbol{\mu}_p^{(s_t)}$. Meanwhile, the state s_t sampled at time t drives the choice of the appropriate parameters $\mathbf{B}_p^{(s_t)}, \boldsymbol{\Psi}_p^{(s_t)}$.

The specification of parameters constrains the OU process to bias the random walk locally, that is in proximity of the patch located at $\boldsymbol{\mu}_p^{(1)}$; alternatively, $\boldsymbol{\mu}_p^{(0)}$ denotes a patch different from current location, which can be reached through displacements at a larger scale defined by $\mathbf{B}_p^{(0)}, \boldsymbol{\Psi}_p^{(0)}$. This way gaze dynamics is given by the multi-state OU equation

$$d\mathbf{r}_F(t) = \mathbf{B}_p^{(s_t)}[\boldsymbol{\mu}_p^{(s_t)} - \mathbf{r}_F(t)]dt + \mathbf{D}_p^{(s_t)}(\mathbf{r}_F(t))d\mathbf{W}^{(s_t)}(t), \quad (18)$$

which is solved by

$$\mathbf{r}(t') \mid \mathbf{r}(t) \sim \mathcal{N}(\boldsymbol{\mu}_p^{(s_t)} + e^{-\mathbf{B}_p^{(s_t)}\delta t}(\mathbf{r}(t) - \boldsymbol{\mu}_p^{(s_t)}), \boldsymbol{\Psi}_p^{(s_t)}). \quad (19)$$

with $\boldsymbol{\Psi}_p^{(s_t)} = \boldsymbol{\Gamma}_p^{(s_t)} - e^{-\mathbf{B}_p^{(s_t)}\delta t}\boldsymbol{\Gamma}_p^{(s_t)}e^{-\mathbf{B}_p^{(s_t)'}\delta t}$. To sum up, gaze dynamics is obtained through the following steps:

1. Sample the behavioural state, based on the current experience of the forager (up to time t, and summarised by parameters $\xi(t)$)

$$s(t) \sim P(\xi(t)) \quad (20)$$

2. Sample the patch index

$$p(t) \sim P(\boldsymbol{\pi}(t)) \quad (21)$$

with $\boldsymbol{\pi}(t)$ the set of parameters depending on the landscape state, and choose patch $\mathcal{P}_p^{(\ell)}$.

3. Set OU parameters $\boldsymbol{\mu}_p^{(s_t)}, \mathbf{B}_p^{(s_t)}, \boldsymbol{\Psi}_p^{(s_t)}$ and sample the gaze shift $\mathbf{r}_F(t) \rightarrow \mathbf{r}_F(t')$ via the OU process specified by Eq. 19, which is explicitly written as

$$\begin{aligned} x_F(t') \mid x_F(t) &\sim \mathcal{N}(\mu_{x,p}^{(s_t)} + e^{-b_{x,p}^{(s_t)}\delta t}(x_F(t) - \mu_{x,p}^{(s_t)}), \psi_{p,x}^{(s_t)}), \\ y_F(t') \mid y_F(t) &\sim \mathcal{N}(\mu_{y,p}^{(s_t)} + e^{-b_{x,p}^{(s_t)}\delta t}(y_F(t) - \mu_{y,p}^{(s_t)}), \psi_{p,y}^{(s_t)}), \end{aligned} \quad (22)$$

with $\psi_{p,x}^{(s_t)} = \gamma_x^{(s_t)}(1 - e^{-2\, b_{x,p}^{(s_t)}\delta t})$ and $\psi_{p,y}^{(s_t)} = \gamma_y^{(s_t)}(1 - e^{-2\, b_{y,p}^{(s_t)}\delta t})$.

Regarding the OU parameters the drift terms $b_{x,p}^{(s_t)}$ and $b_{y,p}^{(s_t)}$ are set proportional to the width of the patch p if $s_t = 1$, or proportional to the distance to the arriving patch (d_p), otherwise. The diffusion terms $\gamma_x^{(s_t)}, \gamma_y^{(s_t)}$ is set proportional to the average distance between patches if $s_t = 0$; equal to 1 otherwise.

Steps 1 and 2 instantiate the choice of the forager's action $\mathcal{A}(t) = \{s(t), p(t)\}$ at time t and involve explicit calculation of Eqs. 20 and 21. These are discussed in the following Section.

B. SWITCHING BEHAVIOUR: SHOULD I STAY OR SHOULD I GO?

Assume that the FoA is located at $\mathbf{r}_F(t)$, within the current patch p, and, for simplicity, that gaze is involved in local patch exploitation. The problem that the perceiver moment to moment has to solve boils down to answering the question: Should I stay or should I go?

In its essence, this is a foraging problem. Indeed, answering such question has long been a fundamental objective in ecology in the endeavour of understanding how animals effectively search for and exploit food patches [96], and, in particular, how a patch cycle is handled. Consider the environment consisting of a set of discrete patches: a cycle starts when the animal leaves a patch to search for a new one; once a patch has been found, the animal gains energy at a rate

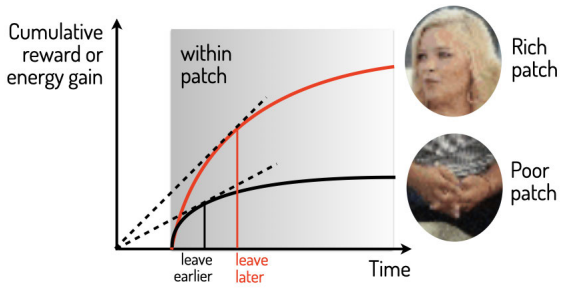


FIGURE 5. The prediction by MVT is that a poor patch should be abandoned earlier than a rich patch. The time axis starts with a travel time with no energy gain after which the forager finds a patch. The shapes of the red and black gain curves, arising from resource exploitation, represent the cumulative rewards of a "rich" and a "poor" patch, respectively. For each curve, the osculation point of the tangent defines the optimal patch residence time.

that decreases as the food becomes depleted; eventually the animal leaves the patch and a new cycle starts.

A series of optimal foraging theories have been developed in line with this objective (see Stephens [13], for a review). By assuming that animal activities are optimized to maximize the rate of net energy gain, optimal foraging theories provide testable hypotheses as well as bases for interpreting complex animal behaviour.

Charnov's marginal value theorem (MVT) is central to these theories [97]. The MVT proposes that foragers should exploit patches in such a way as to maximize a net rate of energy gain and predicts the optimal patch residence time. Let G denote the net energy gain on a cycle, and let T denote the time taken to complete a cycle. Simply put, the MVT states that foragers should move from one patch to another when the marginal rate of food intake (thus, of energy gain, $\partial G/\partial T$) drops to the long-term, average rate $\bar{E}$ of food gain across many patches in the environment.

In this simple model, energy gain is a proxy for fitness and it assumes that the foragers have knowledge about the environment: namely, the quality of other patches and traveling time between patches. Thus, MVT predicts that patch quality should affect patch leaving. Accordingly, a poor patch, yielding a lower energy gain, should be abandoned earlier. Clearly, a forager that stays in a patch too long pays an opportunity cost because it wastes time exploiting a depleted patch when fresher patches remain unexploited.

In a stochastic environment, such as that we are dealing with, where rewards are not deterministic and do not arrive in a smooth flow, an optimal forager should reason about the foraging task probabilistically, based on the potential value of the patch with respect to the environment [79]. The optimal leaving time is when the expected rate, not the observed rate, drops below the average for the environment.

In stochastic foraging models, typically G and T are random variables whose distribution depends on the behavioural strategy adopted by the foraging animal. In particular, G is a function of the time varying state $U(t)$ experienced by the forager up to time t, $G(U(t))$; for instance, as detailed later, the value $U(t) = u$, might indicate the number k of items/preys "captured" by the forager. The mean net rate of energetic gain, or mean reward rate, achieved by the animal is defined as ratio of expectations $\mathbb{E}[G]/\mathbb{E}[T]$.

In a stochastic perspective, it is convenient to consider the instantaneous reward rate [79]

$$g(u,t) = \lim_{\delta t \to 0} \frac{\mathbb{E}[G(U(t+\delta t)) \mid U(t) = u] - G(u)}{\delta t}, \quad (23)$$

that is the expected reward over the next interval of time δt; such definition provides the stochastic counterpart of the continuous energy intake rate $\partial G/\partial T$ exploited by the MVT.

The general rule adopted by the forager, while scrutinising a patch, is to leave the patch when

$$g(u,t) \leq Q(t), \quad (24)$$

that is when the instantaneous reward rate drops below a "quality" threshold Q, which, in general, depends on the richness of the environment, the distance between patches and possibly other factors (in actual foraging, predation risk, etc.).

There is a number of ways to make concrete the rule given in Eq. 24. A method for calculating $g(u,t)$ has been given in Bayesian foraging approaches, e.g. [98], [99].

A straightforward method is the following. Assume that one patch contains a discrete number of items, say m. Let n be the items "consumed" in the time t. Then, the experiential state U is represented by the pair (n,t), $G(U(t)) = G(n,t)$ and $g(u,t) = g(n,t)$. At time t_{W_p} spent within the patch, $k = m-n$ are the items remaining. When foragers search for food items at random, the time required to find one item is assumed to follow the exponential distribution

$$P(T \in [t, t+\delta t]) = \lambda e^{-\lambda t} dt = Ake^{-Akt} dt, \quad (25)$$

where the rate $\lambda = Ak$ depends on A, the searching efficiency of the forager. The probability of capturing at least one item, conditionally on the k remaining, is $P(\delta t \mid k) = 1 - e^{-Akt}$.

It has been calculated [98], [99] that, if the initial distribution of the m_p items in patch p (prior, with $k = m_p$) follows a Poisson law, $Pois(\rho_p) = \frac{e^{-\rho_p}\rho_p^{m_p}}{m_p!}$, then simply

$$g_p(t_{W_p}) = \rho_p e^{-At_{W_p}}. \quad (26)$$

It can be seen from Eq. 26 and Eq. 25 that the foraging efficiency parameter A controls the rate at which the forager switches from one item to another and consequently the instantaneous intake rate. Yet, it is known that individuals concentrate their foraging effort in areas with high reward [100], increasing the handling time of each item, thus increasing the expected time to next item within the patch. In our case, this effect is accounted for by setting $A = \frac{\phi}{\nu_p(t)}$, recalling that $\nu_p(t) \in [0,1]$ is the value associated to the patch p at time t, while ϕ is a positive constant defining the baseline foraging efficiency.

Also, we set ρ as a function of the patch quality, namely,

$$\rho_p(t) = \nu_p(t)|\mathcal{P}_p|e^{-\kappa d_p}, \quad (27)$$

where $|\mathcal{P}_p|$, is the area of the patch, v_p is the patch value, and their product is weighted by $e^{-\kappa d_p}$ representing the visibility of the patch, d_p being the distance to patch p from the current point of gaze, and κ being a positive constant. In foraging terms, the weighting factor accounts for the cost of relocating between patches in foraging.

The expected average gain from the environment for all patches q except the current one can be obtained by considering the potential intake rate at $t_W = 0$, i.e., via Eq. 26 $g_q(0) = \rho_q, q \neq p$:

$$Q(t) = \frac{1}{N_p - 1} \sum_{q \neq p} \rho_q(t). \tag{28}$$

Rather than straightforwardly use the deterministic rule given in Eq. 24, we allow the forager to perform a probabilistic decision; namely the behavioural state decision $s(t) \in 0, 1$ is sampled following a Bernoulli law, $Bern(s(t) \mid \xi(t))$. The parameter ξ, denoting the prior probability of staying within the patch is obtained using a logistic rule accounting for a stochastic comparison on the difference $g_p(t_{W_p}) - Q$, thus

$$\xi(t) = P(stay \mid g(t), Q(t))) = \frac{1}{1 + e^{-\beta(g_p(t_{W_p}) - Q(t))}}, \tag{29}$$

$$s(t) \sim Bern(\xi(t)). \tag{30}$$

By random sampling the behavioural state $s(t)$, most of the time we are likely to get a state "sample" that is somewhere close to the prior $\xi(t)$. However, sometimes we will randomly sample a decision in the tails of the distribution, which offers an opportunity to the forager to tradeoff between the determinism/trend set by rule given Eq. 24, and the dynamically varying landscape.

Eventually, if $s(t) = 0$ is sampled, the choice of a patch is the next step to be accomplished.

C. CHOOSING THE NEXT PATCH

Given the N_P patches, denote π_p the probability of choosing, indexed by patch $p = 1, \cdots, N_P$, with $\sum_p \pi_p = 1$. Then, the sample space of multiple choices can be considered to be the set of 1-of-K encoded random vectors $\mathbf{c}$ of dimension $K = N_P$ having the property that exactly one element c_p has the value 1 and the others have the value 0. The particular element having the value 1 indicates which patch has been chosen. In other terms, $\mathbf{c}$, follows a categorical (or generalised Bernoulli) distribution, $\mathbf{c} \sim Cat(\boldsymbol{\pi}, N_P) = \prod_{p=1}^{N_p} \pi_p^{y_p}$. Probabilities $\boldsymbol{\pi} = (\pi_1 \cdots \pi_{N_P})$ can be related to the above described patch model as follows.

The N_P patches can be considered at time t as sources of independent Poisson processes $M_p(t) \sim Pois(\rho_p(t))$ with mean value function $\mathbb{E}\left[M_p(t)\right] = \rho_p(t)$. Then, in virtue of the superposition theorem [101], the process $M(t) = \sum_{p=1}^{N_p} M_p(t)$ is a Poisson process with expected value $\mathbb{E}[M(t)] = \sum_{p=1}^{N_p} \rho_p(t) = \rho(t)$.

Under such conditions, the coloring theorem holds [102], and the vector $(M_1(t)/S, \ldots, M_{N_p}/S)$, where $S = M_1(t) + \ldots + M_{N_p}$, follows a multinomial distribution with parameters $\pi_p = \frac{\rho_p(t)}{\rho(t)}$.

When considering a single draw, the multinomial distribution is nothing but the categorical distribution; thus, patch choice can be performed by sampling, at any time t the choice vector

$$\mathbf{c} \sim Cat(\pi_1, \cdots, \pi_{N_P}) = \prod_{p=1}^{N_p} \left[\frac{\rho_p(t)}{\rho(t)} \right]^{c_p}, \tag{31}$$

and by selecting patch $\mathcal{P}_p$ based on index p such that $c_p = 1$.

Eq. 31 together with Eqs. 29, 30 completely specify Eqs. 21 and 20, respectively.

VI. SIMULATIONS AND RESULTS

A. METHODOLOGICAL FOREWORD

The rationale behind experiments is to figure out whether simulated behaviours are characterised by statistical properties that are significantly close to those featured by human subjects who have been eye-tracked while watching conversational videos. In simple terms, any model can be considered adequate if model-generated scan paths could have been generated by human observers (which we regard as samples of the `Real` model) while attending to the same audio-visual stimuli.

Consider for example Fig. 6. It summarises the essential spatio-temporal features computed from scan paths that have been sampled via the GazeDeploy procedure (Algorithm 1) on one clip; these are compared to those of human observers on the same clip. Notably, such results are by and large representative of those obtained on the whole dataset.

The simulation has generated scan paths that *prima facie* mimick human scan paths in terms of spatio-temporal statistics. The actual saccade amplitude distribution exhibits a multi modal shape, which is well replicated by the saccades distribution obtained from model simulation (Figure 6b). The model correctly favors small gaze shifts over large ones, that are occasionally undertaken, as highlighted by the right-skewed, long-tailed shape [103]. For what concerns the fixation duration (Fig. 6e), again, distributions from both real and simulated data exhibit a right-skewed and heavy-tailed shape. This is important, since in our model duration is closely related to the modelling of patch giving up time. Apparently, a high similarity can be noticed between saccades direction distributions of real (Fig. 6c) and simulated data (Fig. 6f).

Clearly, beyond the adequate behaviour of the model discernible from such qualitative results, the latter need to be quantitatively substantiated. Are such similarities significant from a statistical standpoint? Is the audio-visual information effectively exploited? Could a different gaze control algorithm provide comparable or even better results?

There are two critical aspects in answering such question.

The first relates to method comparison. Unfortunately a handful of models have been proposed and are experimentally ready for use (i.e., with released code) for predicting gaze shift dynamics. They are referred to as saccadic

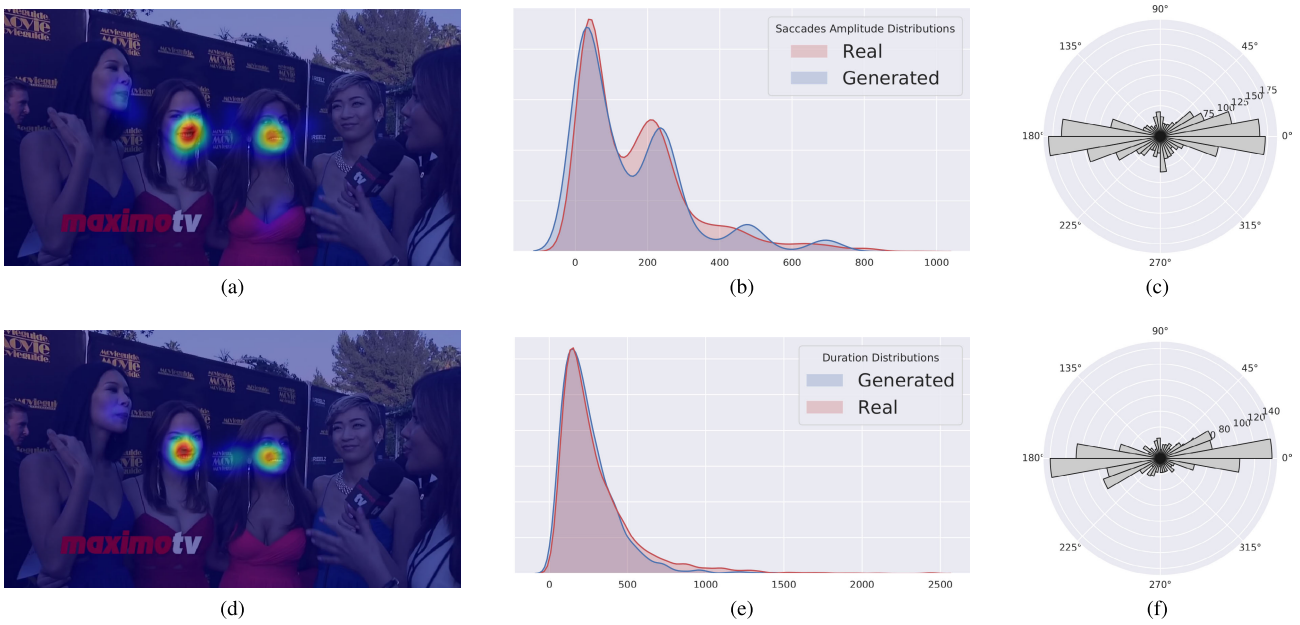


FIGURE 6. **(a) Frame of video 010 with overlaid heatmap of real fixations. (b) Real (red) and Generated (blue) saccades amplitude distribution. (c) Real saccades direction distribution. (d) Frame of video 010 with overlaid heatmap of generated fixations. (e) Real (red) and Generated (blue) fixations duration distribution. (f) Generated saccades direction distribution.**

models [104] and mostly conceived for processing static image input [27], [54], [104]–[107]. Two methods are actually available for handling time-varying stimuli, which we used in our experiment [56], [108].

The second aspect relates to the evaluation metrics. Unlike to classic work on saliency estimation, where standard metrics are available and widely adopted, here assessment must necessarily involve scan path evaluation. There is a lack of consensus about the most appropriate evaluation metrics [109], [110]. In recent years, a number of measures have been proposed, able to deal with the many hurdles of scan path similarity (but for an in-depth review and discussion see [110]). In this work we adopt two well known and state of the art methods: the ScanMatch [111] and the MultiMatch [112], [113] metrics. ScanMatch is apt to provide an overall performance summary, whilst MultiMatch specifically addresses the many dimensions of gaze dynamics. The evaluation of metric results is subtle, thus we support it by addressing appropriate statistical analyses, a point that is often neglected in computational modelling of visual attention.

In this perspective, we switch to a larger dataset - with respect to preliminary experiments reported in [73]-, in terms of conversational episodes, number of participants in the scene, and number of eye-tracked subjects.

B. STIMULI AND EYE-TRACKING DATA

The adopted dataset [17] consists of 65 one-shot conversation scenes from YouTube and Youku, involving 1 to 27 different faces for each scene. The duration of the videos is cut down to be around 20 seconds, with a resolution of 1280×720 pixels at a frame rate of 25 fps. The dataset includes eye-tracking recordings from 39 different participants (26 males and 13 females, ageing from 20 to 49), who were not aware of the purpose of the experiment. The eye fixations position and duration of the 39 subjects were recorded by a Tobii X2-60 eye tracker at 60 Hz.

Ten subjects were randomly sampled out of the 39 and their scan paths used to determine the free parameters of the model described in Section V-B, namely the baseline foraging efficiency ϕ, the logistic growth rate β and the steepness of the exponential determining the visibility of patches κ. A grid search maximising metric scores according to the procedure described in the following Section VI-C yielded as optimal values: $\phi = 3.5$, $\beta = 20$ and $\kappa = 18$.

The remaining 29 subjects were used for evaluation.

C. EVALUATION PROTOCOL

We compare the scan paths simulated from a number of model-based, ''artificial'' observers to those recorded from human observers. By considering different models, or variants of the same model, we simulate different groups of observers. We address two experiments. The first (Sec. VI-D) evaluates the behaviour of the GazeDeploy procedure (thus, exploiting the gaze control strategy described in Algorithm 1) by inhibiting modules accounting for different levels of preattentive information. This provides a family of models, that are ablated variants of what we name the `Full` model.

The second experiment (Sec. VI-E) compares the `Full` model with other gaze control strategies.

In both experiments, the evaluation protocol is the following. For each video:

1) Compute MultiMatch and ScanMatch scores for each possible pair of the 29 real observers (`Real` vs. `Real`).
2) For each model:
   a) Generate gaze trajectories from artificial observers.
   b) Parse/classify trajectories into scan paths (saccades and fixations with the relative duration) via the NSLR-HMM algorithm [114].
   c) Compute MultiMatch and ScanMatch scores for each possible pair of real and 29 artificial scan paths (`Real` vs. `Model`).
3) Return the average ScanMatch and MultiMatch scores for `Real` vs. `Real` and `Real` vs. `Model` comparisons.

As to point 2b), note that (cfr. Fig. 3) the gaze position sequence sampled by GazeDeploy (and its variants) can be assimilated to gaze *raw data* (continuous gaze trajectories) generated by eye-trackers. Thus, in order to follow a classic eye tracking analysis pipeline, the first step is to apply an *event detection* algorithm to both simulated and actual gaze trajectories so to derive the corresponding scan paths (a sequence of fixations). We rely on the NSLR-HMM algorithm described in [114].

For what concerns the metrics, ScanMatch divides a scan path spatially and temporally into several bins and then codes it to form a sequence of letters. Two scan paths are thus encoded as two strings to be compared by maximising the similarity score. This metric indicates the joint spatial, temporal and sequential similarity between two scan paths, higher ScanMatch score denoting a better matching. Complementary, MultiMatch (MM) metrics computes five distinct measures that capture the different scan path features: shape, direction, length, position, and duration. Higher score of each metric means better matching.

In what follows we treat each MultiMatch dimension as a stand-alone score. Thus, the analysis uses six different scores: the five obtained from the MultiMatch (MM) dimensions of shape (MM_{Shape}), direction (MM_{Dir}), length (MM_{Len}), position (MM_{Pos}) and duration (MM_{Dur}), plus the ScanMatch score SM.

D. INFORMATION LEVEL EFFECTS: THE MODEL UNDER THE KNIFE

A basic assumption of the proposed model (A3, Section III), states that in a scene displaying conversations and social interactions, attention is predominantly allocated to faces, with higher relevance given to speakers.

If such premise holds, we expect that the ''ablation'' of model components accounting for face information and specifically for speaker information would lead the model-generated scan paths to significantly deviate, in a statistical sense, from human scan paths.

On the other hand, given that the availability of such information is necessary for a human-like gaze deployment, is it sufficient? To put the question straight: when attending a conversational clip, do we actually need bottom-up information/saliency for reliably generating gaze shifts, or is it redundant? This is a deceptively simple point that has been overlooked, since by and large visual attention models give for granted a central role for low-level salience.

In order to shed light on such questions we simulate gaze data from the following models:

1) `BU` or `Bottom-up`: we prevent the model from the computation of the audio-visual priority maps, thus only considering low-level features in the preattentive stage;
2) `BU+F` or `No Speaker`: faces are considered, together with `BU` features, but no distinction is made between speakers and non-speakers;
3) `F` or `Face` model: only faces are considered, as in the `No Speaker` model, but without `BU` features;
4) `F+S` model: a face and speaker model, without `BU` features;
5) `BU+F+S` or `Full`: the model described in this article where audio-visual patches account for low-level information (`BU`), faces (`F`) and speakers (`S`).

In addition, a baseline `Random` model is adopted, too. This simply generates random gaze shifts by sampling (x, y) fixation coordinates and relative duration from the uniform distribution. Note that in such setting, only the `Full` and the `F+S` models are explicitly accounting for audio information.

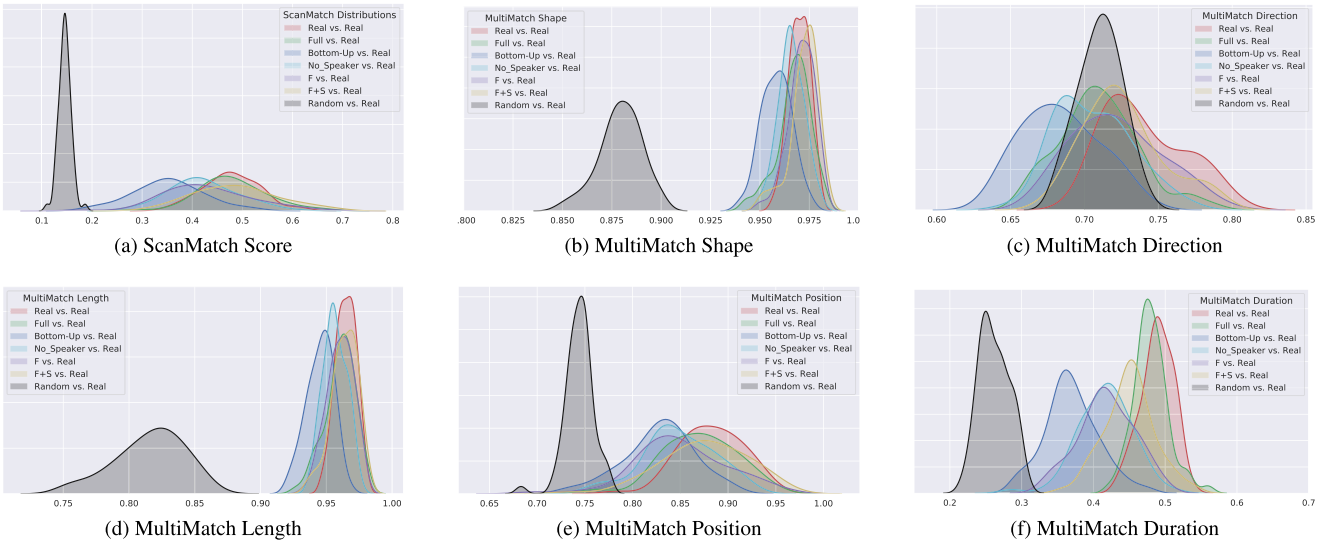
For each model we adopt the protocol described in Section VI-C. Figure 7, depicts at a glance the empirical distributions of the scores obtained in the ablation experiments. A preliminary inspection shows that the `Full` and `F+S` models give rise to distributions that are close to those yielded by real subjects for all dimensions, with the exception of the direction score MM_{Dir}.

1) STATISTICAL ANALYSES

The similarity scores obtained from the six models introduced above are used to assess whether or not a model generates scan paths that significantly differ from those of human observers and to gauge the size of such difference (effect size). In the analyses that follow, scores obtained from `Real` vs. `Real` comparison represent the gold standard; the significance level of all statistical tests is $\alpha = 0.05$.

For each score, the normality of model distributions (groups) was assessed via the Shapiro-Wilk test for normality with Bonferroni correction. All models exhibit normal distributions for scores SM, MM_{Len} and MM_{Dur}; when MM_{Shape}, MM_{Dir} and MM_{Pos} scores were considered, the null hypothesis of normality was rejected for at least one of the models.

Then, for normally distributed scores the statistics adopted to summarise each model were the empirical mean and standard deviation. The effect size for each model was measured via Cohen's d [115], based on differences between model and `Real` means. Otherwise, we considered the median for capturing the central tendency and the absolute deviation from the median as the dispersion measure. In that case



(a) ScanMatch Score (b) MultiMatch Shape (c) MultiMatch Direction

(d) MultiMatch Length (e) MultiMatch Position (f) MultiMatch Duration

FIGURE 7. Score distributions for models considered in the ablation experiment.

TABLE 2. Information level effects: central tendencies for each score and model computed as mean (M) or median (MED) with associated dispersion metrics (standard deviation, SD or median absolute deviation, MAD. Effect sizes are computed as the Cohen's d or the Cliff's δ between the given model and real subjects.

	M	SD	d	Magnitude
F+S	0.487	0.082	-0.127	negligible
Full	0.481	0.067	-0.045	negligible
Face	0.430	0.088	0.650	medium
No Speaker	0.423	0.067	0.889	large
Bottom-Up	0.352	0.071	1.955	large
Random	0.146	0.012	8.140	large
Real	0.478	0.056	0	/

(a) ScanMatch Score

	MED	MAD	δ	Magnitude
F+S	0.974	0.006	-0.300	small
Face	0.971	0.007	-0.140	negligible
Full	0.969	0.007	0.170	small
No Speaker	0.965	0.006	0.391	medium
Bottom-Up	0.959	0.009	0.790	large
Random	0.880	0.010	1.000	large
Real	0.970	0.005	0	/

(b) MultiMatch Shape

	MED	MAD	δ	Magnitude
F+S	0.722	0.022	0.283	small
Face	0.718	0.030	0.351	medium
Random	0.711	0.016	0.651	large
Full	0.707	0.024	0.574	large
No Speaker	0.708	0.030	0.628	large
Bottom-Up	0.680	0.024	0.862	large
Real	0.734	0.029	0	/

(c) MultiMatch Direction

	M	SD	d	Magnitude
F+S	0.964	0.010	0.116	negligible
Full	0.960	0.011	0.462	small
Face	0.961	0.010	0.486	small
No_Speaker	0.957	0.009	1.034	large
Bottom-Up	0.945	0.010	2.186	large
Random	0.816	0.026	7.726	large
Real	0.965	0.007	0	/

(d) MultiMatch Length

	MED	MAD	δ	Magnitude
F+S	0.878	0.048	0.093	negligible
Full	0.869	0.044	0.205	small
Face	0.841	0.040	0.459	medium
No Speaker	0.844	0.035	0.516	large
Bottom-Up	0.830	0.034	0.748	large
Random	0.745	0.012	1.000	large
Real	0.885	0.038	0	/

(e) MultiMatch Position

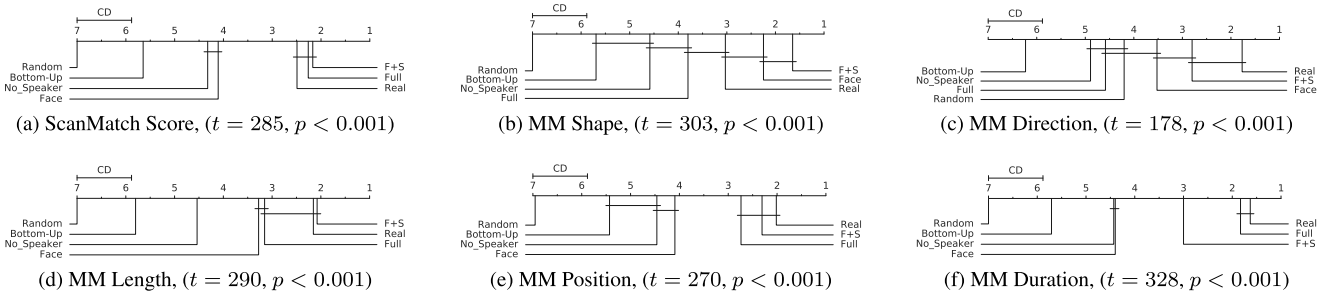
	M	SD	d	Magnitude
Full	0.480	0.024	0.384	small
F+S	0.450	0.035	1.328	large
Face	0.418	0.038	2.269	large
No Speaker	0.416	0.037	2.355	large
Bottom-Up	0.370	0.037	3.866	large
Random	0.262	0.021	10.603	large
Real	0.489	0.022	0	/

(f) MultiMatch Duration

the effect size for each model was computed via Cliff's delta [116].

The overall results are reported in Table 2. We follow Cohen's convention [115] considering effect magnitudes 'small' ($d \sim 0.2$), 'medium' ($d \sim 0.5$), 'large' ($d \sim 0.8$) and negligible ($d < 0.2$). As to Cliff's delta, we follow Hess and Kromrey [117], by distinguishing 'small' ($\delta \sim 0.147$), 'medium' ($\delta \sim 0.33$) and 'large' ($\delta \sim 0.474$) effect size; the effect is negligible for $\delta < 0.147$.

We then performed homogeneity of variance tests. For each score, when normality held, the Bartlett's test was employed to test homoscedasticity; otherwise, Levene's test



(a) ScanMatch Score, ($t = 285$, $p < 0.001$) (b) MM Shape, ($t = 303$, $p < 0.001$) (c) MM Direction, ($t = 178$, $p < 0.001$)

(d) MM Length, ($t = 290$, $p < 0.001$) (e) MM Position, ($t = 270$, $p < 0.001$) (f) MM Duration, ($t = 328$, $p < 0.001$)

FIGURE 8. Information level effects: critical Difference (CD) diagrams of the post-hoc Nemenyi test ($\alpha = 0.05$) for the ScanMatch score and each MultiMatch score obtained by using different information levels obtained by ablation of components feeding the GazeDeploy strategy. Diagrams can be read as follows: the difference between two models is significant if the gap between their ranks is larger than CD; there is a line between two models if the rank gap between them is smaller than CD. Graphically, models that are not significantly different from one another are connected by a black CD line. Friedman's test statistic (*t*) and p-value (*p*) are reported in brackets.

was adopted. Either Bartlett's or Levene's tests rejected the null hypothesis of homogeneity of variances ($p < 0.01$, for all scores).

The assessment of statistically significant differences between models was performed as follows. Since neither normality, nor equality of variances could be ensured, we resorted to the well known Friedman Test (FT, [118], a non-parametric variant of ANOVA), with Nemenyi [119] *post-hoc* analysis of pairwise differences (similar to the Tukey test for ANOVA). We tested the null hypothesis for each score that the medians were equal between the 6 groups plus the `Random` one.

For all scores, the FT rejected the null hypothesis ($p < 0.001$, always, cfr. Fig. 8). Thus, for each score at least one statistically significant difference between two models exists. The Nemenyi's *post-hoc* analysis was then performed. The test compares each pair of groups in terms of their difference in average ranks; if such difference exceeds the critical difference CD_α at the confidence level α, then the two group are statistically different. Figure 8 reports the FT outcomes (test statistics t and p-value p) and, most important, visualises *post-hoc* analysis results. The latter are rendered in a compact, information-dense format by means of the *Critical Difference (CD) Diagram* as proposed in [120]. CD Diagrams show the average rank of each model (higher ranks meaning higher average scores); models whose difference in ranks does not exceed the CD_α ($\alpha = 0.05$) are joined by thick lines and cannot be considered significantly different.

By first considering the ScanMatch metric, a clear ranking is established. We can assume that there are no significant differences within the following two groups: `F+S`, `Real` and `Full`; `Face` and `No_Speaker`. All other differences are significant. Taking into account the magnitude of the effect, the difference between the `Full` model and `Real` is negligible with a smaller magnitude than that of `F+S`. The effect size grows to large for `Face` and `No_Speaker`. The `Bottom-Up` model performs badly, albeit being significantly different from the `Random` model, which clearly has the largest effect size.

Together with the fact that when the `BU` component is ablated from higher level models, the similarity performance does not decrease, these results suggest that `BU` conspicuity has a modest relevance, at least for the kind of conversational clips we deal with.

Overall, it can be noted that the test does not support any statistically significant difference between the scores of `Real` subjects and the ones from the `Full` model. This is true for the ScanMatch metric and for all the MultiMatch metric dimensions except for the *Direction* score. A similar behaviour is exhibited by the `F+S` model, the only remarkable difference being that of the MultiMatch *Duration* metric. In this case, as opposed to the `Full` model a significant difference with fixation duration of real subjects is found. Significant differences arise when comparing real scan paths with those generated from ablated models like the `Bottom-Up`, `No Speaker` and `Face` and the huge dissimilarity with the randomly generated eye movements (`Random` model).

Taken together with the size effects reported in Table 2, these results bear some consequences. First of all, they show how the proposed (`Full`) model is able to mimic the human behaviour of gaze deployment to audio-visual dynamic stimuli of social interactions. This is witnessed by differences with the scores achieved by real subjects that are negligible in their size and not statistically significant for almost all scores. The only exception is the MultiMatch *Direction* dimension, for which no clear association with the *gold standard* is found. This is not surprising. Indeed the saccades direction is the only feature that is not explicitly tackled in any aspect of the proposed model, but only subsumed as a consequence of the value based patch selection mechanism (Eq. 31).

Second, it is interesting to note how preventing models from accounting for bottom-up information, does not results in a significant loss of performance, according to most of the adopted metrics, when comparing with the same models that account for it. Indeed, if fixation duration seems to benefit from the computation of low level cues, for other scores like the *ScanMatch* and the MultiMatch *Position*, the ablation of bottom-up information generated outcomes that are

indistinguishable from a statistical standpoint. In light of this result, it is clear how the role of bottom-up information when dealing with videos of social interaction, should be reappraised, since marginally contributing to the process of attention allocation.

Overall, the only model that performs comparably with humans is the `Full` model; indeed it is able to achieve indistinguishable results w.r.t. humans on 5 out of the 6 adopted metrics. The fact that the models obtained after the ablation of high level information (speaker/no-speaker, face location) produce significantly lower scores, highlights the causal effect of the presence of (talking) faces, or more generally top down cues, on attention allocation. This fact has been previously demonstrated in the psychological field [12], but here it is made operational by means of a computational model.

E. GAZE CONTROL EFFECTS

The esperiment reported in this section aimed at comparing the GazeDeploy control strategy to those of models previously proposed in the literature that are 1) capable of handling time-varying scenes and for which 2) a model implementation is available. In particular we used the Ecological sampling model (from now on `Eco_Sampling`)[2] proposed in [56], and the recent `G-Eymol` model[3] [108].

In a nutshell, `Eco_Sampling` is a stochastic model of eye guidance, much like GazeDeploy. The gaze shift dynamics is implemented in terms of a stochastic differential equation driven by α-stable noise, and grounds its motivation in the Lévy flight approaches to foraging displacements [15], [121]. Different from GazeDeploy it does not rely on a specific account for patch handling and giving-up time. The preattentive representation is formalised in terms of proto-objects, roughly corresponding to patches. The overall control strategy is based on a complexity measure of the perceived time-varying scene. Complexity is computed from interest points that are stochastically sampled from the proto-object representation. Here, we feed the model with the same perception of the world as inferred in the preattentive stage of the proposed `Full` model, so as to focus on the performance of the different gaze control strategies, rather than representation issues.

The `G-Eymol` model generates gaze trajectories via differential equations of motion derived through variational laws somehow related to mechanics. The focus of attention is subject to a gravitational field. The distributed virtual mass that drives eye movements is associated with the presence of details and motion in the video. The inhibition of return (IOR, [122]) mechanism is employed to avoid the model being stuck in the same portions of the visual landscape. Such virtual masses are proportional to the amount of details and motion of the scene, defined as the magnitude of the gradient and the magnitude of the optical flow, respectively. Authors suggest that top-down information can considered by defining object-based gravitational attractors. The original implementation relies on the Haar cascade face detection [123], that allows faces as additional masses. This is the `G-Eymol` version we adopt here. Further, in order to belay a fair comparison, we set up a variant (`G-Eymol_sp`) that takes into account the difference between speakers and non-speakers. This is achieved by feeding the `G-Eymol` model with speaker and non-speaker masses whose magnitude is proportional to their value as defined in Eq. 8. The `G-Eymol` equation of motion are deterministic. However, the stochasticity requested to sample different scan paths mimicking different observers can be achieved by perturbing the initial conditions of the equations. Eventually, we also consider the `Random` model.

As in the previous experiment, for each model we adopted the protocol described in Section VI-C. Figure 9, depicts at a glance the empirical distributions of the scores obtained by the 5 control models. Visual inspection of distributions derived from the ScanMatch score suggests a higher similarity of scan paths simulated from the `Full` model with respect to the original `G-Eymol`; `Eco_Sampling` and, surprisingly, `G-Eymol_sp` achieve inferior performance.

As to MultiMatch, the behaviour of the `Full` model gives rise to distributions that on the average are closer than other models to those yielded by real subjects, the MM direction score again being an exception, as in the previous experiment. But here, remarkably, the `Full` model seems to outperform the others with respect to the Duration dimension, a result that was to be expected, because this dimension benefits from the Bayesian stochastic foraging approach.

1) STATISTICAL ANALYSES

The statistical analysis of effects entailed by different gaze control strategies, closely followed the one carried out in Section VI-D1.

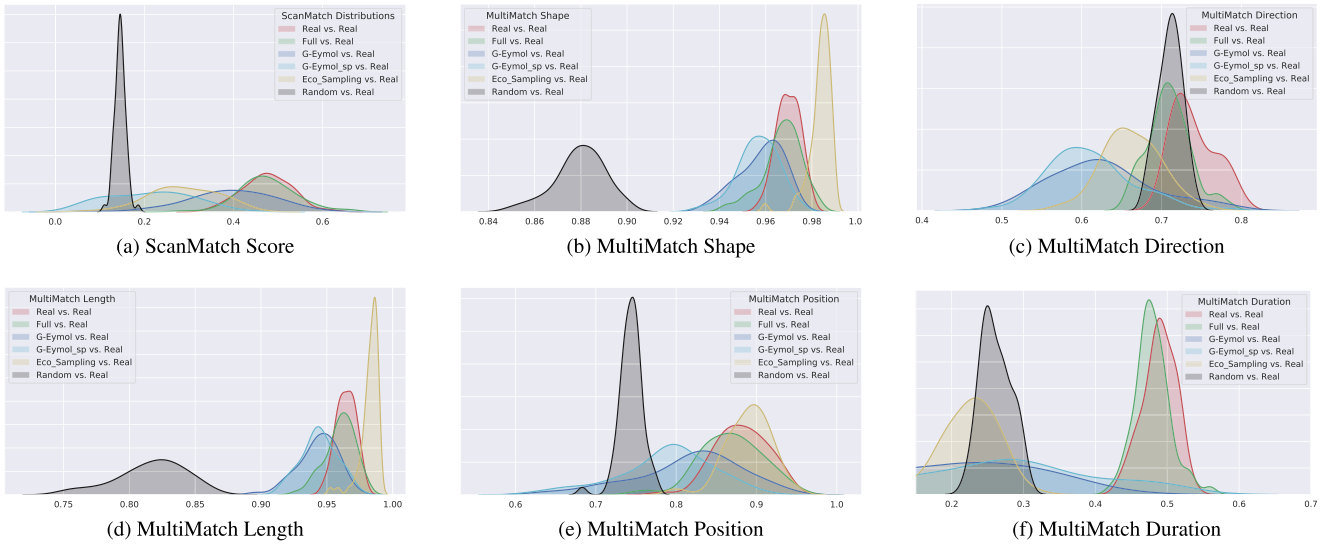
Scores SM, MM_{Dir}, MM_{Pos} and MM_{Dur} according to the Shapiro-Wilk test failed to reject the hypothesis of normality, as opposed to scores MM_{Shape} and MM_{Len}. The overall results for central tendencies, dispersions and effect size are reported in Table 3. For all scores, either Bartlett's or Levene's tests rejected the hypothesis of homoscedasticity of distributions. Thus, the FT with Nemenyi *post-hoc* analysis was performed. The final results are reported in Figure 10.The quantitative results overall support what surmised so far by visually inspecting the score distributions.

Based on the *post-hoc* Nemenyi test, and considering the ScanMatch metric, we assume that there are no significant differences within the following three groups: `Full`, `Real`; `Eco_Sampling` and `G-Eymol_sp`; `G-Eymol_sp` and `Random`. All other differences are significant. The effect size of such differences with respect to the gold standard of human observers can be appreciated in Table 3.

The many facets of such overall model performance ranking can be best weighed by considering the individual

[2]Matlab implementation available at https://github.com/phuselab/EcoSampling

[3]Python implementation available at https://github.com/dariozanca/G-Eymol



(a) ScanMatch Score (b) MultiMatch Shape (c) MultiMatch Direction

(d) MultiMatch Length (e) MultiMatch Position (f) MultiMatch Duration

FIGURE 9. Score distributions for models considered in the gaze control experiment.

TABLE 3. Gaze control effects (notation follows Table 2).

	M	SD	d	Magnitude
Full	0.477	0.065	-0.044	negligible
G-Eymol	0.393	0.098	1.031	large
Eco_Sampling	0.288	0.078	2.779	large
G-Eymol_sp	0.212	0.093	3.441	large
Random	0.146	0.012	8.393	large
Real	0.475	0.054	0	/

(a) ScanMatch Score

	MED	MAD	δ	Magnitude
Eco_Sampling	0.985	0.003	-0.939	large
Full	0.968	0.007	0.171	small
G-Eymol_sp	0.957	0.008	0.820	large
G-Eymol	0.960	0.008	0.714	large
Random	0.881	0.010	1.000	large
Real	0.969	0.006	0	/

(b) MultiMatch Shape

	M	SD	d	Magnitude
Random	0.711	0.015	1.345	large
Full	0.710	0.027	1.131	large
Eco_Sampling	0.663	0.036	2.424	large
G-Eymol	0.626	0.064	2.343	large
G-Eymol_sp	0.610	0.053	3.065	large
Real	0.741	0.027	0	/

(c) MultiMatch Direction

	MED	MAD	δ	Magnitude
Eco_Sampling	0.985	0.004	-0.913	large
Full	0.960	0.011	0.220	small
G-Eymol_sp	0.944	0.013	0.803	large
G-Eymol	0.944	0.012	0.810	large
Random	0.820	0.024	1.000	large
Real	0.964	0.009	0	/

(d) MultiMatch Length

	M	SD	d	Magnitude
Eco_Sampling	0.890	0.028	-0.285	small
Full	0.868	0.039	0.364	small
G-Eymol	0.816	0.066	1.273	large
G-Eymol_sp	0.787	0.059	1.991	large
Random	0.744	0.014	5.549	large
Real	0.881	0.032	0	/

(e) MultiMatch Position

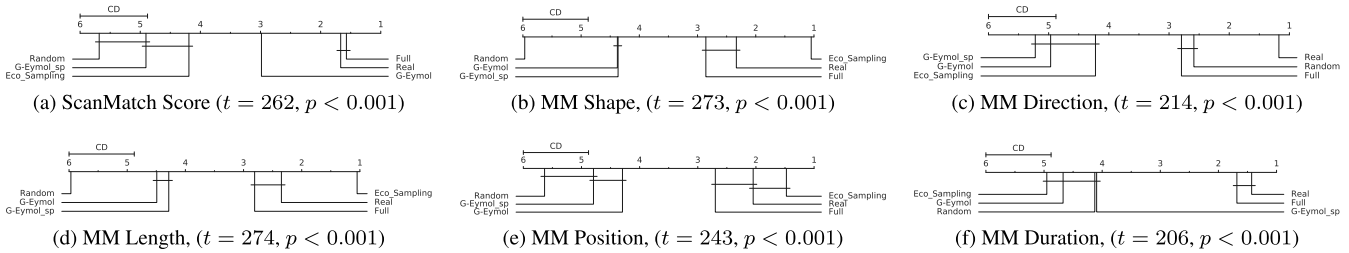
	M	SD	d	Magnitude
Full	0.480	0.024	0.389	small
G-Eymol_sp	0.278	0.122	2.411	large
Random	0.261	0.021	10.562	large
G-Eymol	0.213	0.107	3.556	large
Eco_Sampling	0.221	0.044	7.677	large
Real	0.489	0.022	0	/

(f) MultiMatch Duration

dimensions provided by MultiMatch scores. Remarkable is the divergence with respect to duration: no significant differences are detected within the `Full`and `Real` group, with a small effect magnitude; all other models fall in one group, together with the `Random` model, albeit distinguished by different effect sizes (in this case, for instance `G-Eymol_sp` performs better than the original `G-Eymol`).

The worst performance of all models is detectable on the direction dimension. In this case there are no significant differences within the following two groups: `Full` and `Random` (but with smaller effect size for the first model); `Eco_Sampling`, `G-Eymol` and `G-Eymol_sp`. By taking into account the effect size, models in the second group perform worse than random choice. In simple terms, a random choice of direction seems to provide a better opportunity than any inappropriate strategy.

As to the position dimension there are no significant differences within the following groups: `Eco_Sampling` and `Real`; `Full` and `Real`; `G-Eymol` and `G-Eymol_sp`; `G-Eymol_sp` and `Random`. The smallest effect size in



(a) ScanMatch Score ($t = 262$, $p < 0.001$) (b) MM Shape, ($t = 273$, $p < 0.001$) (c) MM Direction, ($t = 214$, $p < 0.001$)

(d) MM Length, ($t = 274$, $p < 0.001$) (e) MM Position, ($t = 243$, $p < 0.001$) (f) MM Duration, ($t = 206$, $p < 0.001$)

FIGURE 10. Gaze control effects: CD Diagrams of the post-hoc Nemenyi test ($\alpha = 0.05$) for MultiMatch (MM) and ScanMatch scores (cfr. Fig 8), obtained by using different gaze control strategies (see text for explanation). Friedman's test statistic (t) and p-value (p) are reported in brackets.

difference from human subjects are provided by the `Full` and `Eco_Sampling` models, the latter being the smallest. This could be due to the fact that, all being equal, the sampling mechanism of interest points from proto-objects/patches can provide a fine-grained, shape-sensitive choice of the possible gaze shift "landing", as opposed to the use of the center of mass attraction in GazeDeploy. Similarly, the high ranking of `Eco_Sampling` is likely to stem from the core business of the approach, namely the sophisticated modelling of gaze shift amplitudes via Lévy flights; yet, GazeDeploy suffers from a smaller effect size. Results achieved for the shape score deserve similar considerations.

VII. CONCLUSION

In our daily life we orient our attention and move our gaze to gauge and collect information that includes social cues. Conversational videos have the ecological virtue of displaying many such cues embedded in a dynamic, albeit controlled situation. Hence, their analysis brings in fundamental questions on the attentive behaviour of a subject who scrutinises and forages on other subjects involved in social interactions: *What* defines a patch of audio-visual information valuable to spot? *How* is gaze guided within and between patches?

Surprisingly the study of this problem is still in its infancy in the field of computational modelling of visual attention [32], [124], [125]. This state of affairs is in striking contrast with the exponentially spreading body of audio-visual data that convey social content and the need of analysing the perceiver's behaviour under such circumstances. Unwisely, a large amount of research effort in the last two decades has by and large focused on salience estimation from natural scenes, mostly neglecting the dynamics of actual attention deployment, as instantiated by gaze shifts. The shortcomings of this effort become dauntingly palpable when dealing with scenes endowed with rich semantics, where gaze sampling is affected by goals, rewards, personal social traits and even expectations about future events.

Here we deliberately made a fresh step forward in such direction. Gaze dynamics has been derived in a principled way by reformulating attention deployment as a stochastic foraging problem: the perceiver allocates gaze to audio-visual patches much like a forager visits patches in the environment to obtain nourishment. Our model is that of a stochastic forager performing an Ornstein-Uhlenbeck walk by switching to the appropriate scale for engaging in either within-patch exploitation and large between-patch relocations. The foraging dynamics is thus driven by the audio-visual patches that at any time appear relevant as to social value (rewarding). Patches are sampled from spatially-based probabilistic priority maps. These, in turn, are derived by adapting to our framework recent results gained by deep network techniques [126], [127], so to account for the visual and auditory objects across different analysis scales. Moment to moment, patch value dynamics is inferred on a video clip, with dynamics parameters being derived on the basis of eye-tracked gaze allocation of a number of actual observers. Patch choice, handling and leave are framed within an optimal Bayesian foraging setting.

Model simulation experiments on a publicly available dataset of eye-tracked subjects and in-depth statistical analyses of results so far achieved have been performed. These show an overall statistically significant similarity between scan paths of human observers and those generated by the GazeDeploy procedure, which uses the full stack of information levels.

The current model has limitations that pave the way for future research. For instance, statistical analyses have highlighted specific problems in gaze direction modelling. This is a difficult hurdle to face. Some contextual rules (e.g., the prevalence of horizontal scanning) that have been advocated in the computer vision field [128] and in the psychological literature [42], might fail in more ecological conditions, out of the lab and in dynamic environments. On the other hand, the ecology of animal movements is still struggling on the point [15] in spite of an important body of research laid down over years. One solution could be that of a data-driven strategy [45], [50], albeit raising in turn the problem of generalisability.

The model simulates fixation duration from first principles (Charnov's theorem) and achieves significant performance. Notwithstanding, it would be interesting to amend the lack of an explicit account for actual patch exploitation and within-patch item handling. One such example is facial expression processing of people engaged in the conversations. Expression perception is one fundamental mean for our understanding of and engagement in social interactions. This aspect is intimately related to the notion of value proposed in our work, which represents as a matter of fact a doorway to intertwine attention, cognition and emotion [38].

Indeed, several studies have reported the influence of emotion on overt attention and emphasised the distinction between internally and externally located emotional cues; meanwhile, other studies have shown the reversed causal effect: attention can also affect emotional responses [124], [129].

Despite of such limitations, the results of this study allow to draw at least two general conclusions.

The first lies in that when we engage with the computational modelling of attention in multimodal scenarios with rich semantics, we should not overstate the role of classic salience. Concentrating all research efforts by mostly focussing on subtle improvements of such techniques (whose statistical significance is at best questionable), under the wishful assumption that these will be predictive of actual gaze allocation, might not be the optimal strategy.

The second and artful lesson to learn, is that general models of gaze deployment are appealing, indeed elegant and explainable. Nevertheless, they should be general as to the foundational principles and rationales, albeit not generic. Caution suggests that the many dimensions of gaze dynamics are to be specifically accounted for, if the similarity to human gaze behaviour is the ultimate goal.

APPENDIX A
DERIVING FEATURE MAPS

The input stimuli $\mathcal{S}$ are represented by the time-varying visual and audio streams, $\mathcal{S}(t) = \{\mathbf{I}(t), \mathbf{A}(t)\}, t = 1, \cdots, T$, where $\mathbf{I}$ is the frame sequence and $\mathbf{A}$ the audio signal.

In order to derive a priority map, we need to specify which features $\mathbf{F}$ are to be taken into account, given the context or goal $\mathcal{G}$, and the feature maps $\mathbf{X}$, that is the topographically organised maps that encode the joint occurrence of a specific feature at a spatial location [28]. In a probabilistic setting, a feature map $\mathbf{X}_f$ is a matrix of binary RVs $x(\mathbf{r})$ denoting whether feature f is present or not present at location $\mathbf{L} = \mathbf{r}$ [28]. It can be equivalently represented as a unique map encoding the presence of different object dependent features $\mathbf{F}_{f,\mathbf{O}}$, or a set of object-specific feature maps, i.e. $\mathbf{X} = \{\mathbf{X}_f\}$ (e.g., in the visual realm, a face map, a body map, etc.)

2) VISUAL FEATURES

From input $\mathbf{I}$, two kinds of visual features are derived: generic visual features $\mathbf{F}_\mathbf{I}$ - such as edge, texture, colour, motion features-, and object-dependent features, $\mathbf{F}_{\mathbf{O}_V}$. The latter are selected by taking into account the classes of objects that are likely to be relevant under the goal $\mathcal{G}$. Internal goals are biased towards social cues, thus the prominent visual objects are faces, $\mathbf{O}_V = \{face\}$. Both kinds of visual features, $\mathbf{F}_\mathbf{I}$ and $\mathbf{F}_{\mathbf{O}_V}$, can be estimated in a feed-forward way.

Features $\mathbf{F}_\mathbf{I}$ and $\mathbf{F}_{\mathbf{O}_V}$ need to be spatially organised in feature maps. In the visual attention context, the distribution $P(\mathbf{X})$ can be considered the probabilistic counterpart of the classic salience map [28]. Thus, $\mathbf{X}_{f,\mathbf{I}}$ represents the support of a low-level saliency map, whilst $\mathbf{X}_{f,\mathbf{O}_V}$ is the support of an high-level, object-based saliency map. At this stage, the inferential step entails estimating the posteriors $P(\mathbf{X}_\mathbf{I} \mid \mathbf{F}_\mathbf{I})$ and $P(\mathbf{X}_{\mathbf{O}_V} \mid \mathbf{F}_{\mathbf{O}_V})$, whatever the technique adopted.

In order to derive the physical stimulus feature map $\mathbf{X}_\mathbf{I}$, we rely on the spatio-temporal saliency method proposed in [130] based on local regression kernel center/surround features. It avoids specific optical flow processing for motion detection and has the advantage of being insensitive to possible camera motion. By assuming uniform prior on all locations, the evidence from a location $\mathbf{r}$ of the frame is computed via the likelihood $P(\mathbf{I}(t) \mid \mathbf{x}_f(\mathbf{r}, t) = 1, \mathbf{F}_\mathbf{I}, \mathbf{r}_F(t)) = \frac{1}{\sum_s} \exp\left(\frac{1-\rho(\mathbf{F}_{\mathbf{r},c}, \mathbf{F}_{\mathbf{r},s})}{\sigma^2}\right)$, where $\rho(\cdot) \in [-1, 1]$ is the matrix cosine similarity (see [130], for details) between center and surround feature matrices $\mathbf{F}_{\mathbf{r},c}$ and $\mathbf{F}_{\mathbf{r},s}$ computed at location $\mathbf{r}$ of frame $\mathbf{I}(t)$.

The visual object-based feature map $\mathbf{X}_{\mathbf{O}_V}$ entails a face detection step. The method proposed by Hu and Ramanan [131] has shown, in our preliminary experiments, to bear the highest performance. It relies on a feed-forward deep network architecture for scale invariant detection. Starting with an input frame $\mathbf{I}(t)$, a coarse image pyramid (including interpolation) is created. Then, the scaled input is fed into a Convolutional Neural Network (CNN) to predict template responses at every resolution. Non-maximum suppression (NMS) is applied at the original resolution to get the final detection results. Their confidence value is used to assign the probability $P(\mathbf{X}_{\mathbf{O}_V} \mid \mathbf{F}_{\mathbf{O}_V}, \mathbf{L}_V = \mathbf{r})$ of spotting face features $\mathbf{F}_{\mathbf{O}_V}$ at $\mathbf{L}_V = \mathbf{r}$, according to a gaussian distribution located on the face center modulated by detection confidence and face size.

3) AUDIO AND AUDIO-VISUAL FEATURES

From input $\mathbf{A}$, in our setting the objects of interest $\mathbf{O}_A$ are represented by speakers' voices [12], and features $\mathbf{F}_{f,\mathbf{O}_A}$ suitable to represent speech cues. In this work, we are not considering other audio sources (e.g, music). We are interested in inferring the audio-visual topographic maps of speaker/non-speakers, $\mathbf{X}_{\mathbf{O}_{AV}}$, given the available faces in the scene and speech features via the posterior distribution $P(\mathbf{X}_{\mathbf{O}_{AV}} \mid \mathbf{X}_{\mathbf{O}_A}, \mathbf{X}_{\mathbf{O}_V}, \mathbf{F}_{\mathbf{O}_A}, \mathbf{F}_{\mathbf{O}_V})$, where $\mathbf{X}_{\mathbf{O}_{AV}} = x(\mathbf{r})$ denotes whether a speaker/non-speaker is present or not present at location $\mathbf{r}$.

Technically, the features $\mathbf{F}_{\mathbf{O}_A}$ used to encode the speech stream are the Mel-frequency cepstral coefficients (MFCC). The audio feature map $\mathbf{X}_{\mathbf{O}_A}(t)$ can be conceived as a spectro-temporal structure computed from a suitable time window of the audio stream, representing MFCC values for each time step and each Mel frequency band. It is important to note, that the problem of deriving the speaker/non-speaker map $\mathbf{X}_{\mathbf{O}_{AV}}$ when multiple faces are present, is closely related to the AV synchronisation problem [126]; namely, that of inferring the correspondence between the video and the speech streams, captured by the joint probability $P(\mathbf{X}_{\mathbf{O}_{AV}}, \mathbf{X}_{\mathbf{O}_A}, \mathbf{X}_{\mathbf{O}_V}, \mathbf{F}_{\mathbf{O}_A}, \mathbf{F}_{\mathbf{O}_V}, \mathbf{L}_{AV})$. The speaker's face is the one with the highest correlation between the audio and the video feature streams, whilst a non-speaker should have a correlation close to zero. It has been shown that the

synchronisation method presented in [126] can be extended to locate the speaker vs. non-speakers and to provide a suitable confidence value. The method relies on a two-stream CNN architecture (SynchNet) that enables a joint embedding between the sound and the face images. In particular we use the Multi-View version [126], [127]), which allows the speaker identification on profile faces and does not require explicit lip detection. To such end, 13 Mel frequency bands are used at each time step, where features $\mathbf{F}_{\mathbf{O}_A}(t)$ are computed at sampling rate for a 0.2-secs time-window of the input signal $\mathbf{A}(t)$. The same time-window is used for the video stream input.

ACKNOWLEDGMENT

Alessandro D'Amelio thanks Prof. Tom Foulsham for enlightening discussions.

REFERENCES

[1] A. Truong and M. Agrawala, "A tool for navigating and editing 360 video of social conversations into shareable highlights," in *Proc. 45th Graph. Interface Conf. Proc. Graph. Interface 2019*. Toronto, ON, Canada: Canadian Human-Computer Communications Society, 2019, pp. 1–9.

[2] K. Pires and G. Simon, "YouTube live and twitch: A tour of user-generated live streaming systems," in *Proc. 6th ACM Multimedia Syst. Conf. (MMSys)*, 2015, pp. 225–230.

[3] A. Nassauer and N. M. Legewie, "Analyzing 21st century video data on situational dynamics," issues and challenges in video data analysis," *Social Sci.*, vol. 8, no. 3, p. 100, 2019.

[4] J. P. Staab, "The influence of anxiety on ocular motor control and gaze," *Current Opinion Neurol.*, vol. 27, no. 1, pp. 118–124, Feb. 2014.

[5] R. B. Grossman, J. Mertens, and E. Zane, "Perceptions of self and other: Social judgments and gaze patterns to videos of adolescents with and without autism spectrum disorder," *Autism*, vol. 23, no. 4, pp. 846–857, May 2019.

[6] M. Jording, D. Engemann, H. Eckert, G. Bente, and K. Vogeley, "Distinguishing social from private intentions through the passive observation of gaze cues," *Frontiers Hum. Neurosci.*, vol. 13, p. 442, Dec. 2019.

[7] N. Guy, H. Azulay, R. Kardosh, Y. Weiss, R. R. Hassin, S. Israel, and Y. Pertzov, "A novel perceptual trait: Gaze predilection for faces during visual exploration," *Sci. Rep.*, vol. 9, no. 1, pp. 1–12, Dec. 2019.

[8] J. M. Henderson, "Gaze control as prediction," *Trends Cognit. Sci.*, vol. 21, no. 1, pp. 15–23, Jan. 2017.

[9] M. F. Land, "Eye movements and the control of actions in everyday life," *Prog. Retinal Eye Res.*, vol. 25, no. 3, pp. 296–324, May 2006.

[10] S. V. Shepherd and M. L. Platt, "Spontaneous social orienting and gaze following in ringtailed lemurs (lemur catta)," *Animal Cognition*, vol. 11, no. 1, p. 13, May 2007.

[11] E. Kowler, "Eye movements: The past 25 years," *Vis. Res.*, vol. 51, no. 13, pp. 1457–1483, Jul. 2011.

[12] T. Foulsham, J. T. Cheng, J. L. Tracy, J. Henrich, and A. Kingstone, "Gaze allocation in a dynamic situation: Effects of social status and speaking," *Cognition*, vol. 117, no. 3, pp. 319–331, Dec. 2010.

[13] D. W. Stephens, *Foraging Theory*. Princeton, NJ, USA: Princeton Univ. Press, 1986.

[14] F. Bartumeus and J. Catalan, "Optimal search behavior and classic foraging theory," *J. Phys. A, Math. Theor.*, vol. 42, no. 43, Oct. 2009, Art. no. 434002.

[15] G. M. Viswanathan, M. G. Da Luz, E. P. Raposo, and H. E. Stanley, *The Physics of Foraging: An Introduction to Random Searches and Biological Encounters*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[16] J. M. Wolfe, "When is it time to move to the next raspberry bush? Foraging rules in human visual search," *J. Vis.*, vol. 13, no. 3, p. 10, Jan. 2013.

[17] M. Xu, Y. Liu, R. Hu, and F. He, "Find who to look at: Turning from action to saliency," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4529–4544, Sep. 2018.

[18] W. James, *The Principles of Psychology*. New York, NY, USA: Dover, 1890.

[19] A. Yarbus, *Eye Movements and Vision*. New York, NY, USA: Plenum Press, 1967.

[20] B. W. Tatler, M. M. Hayhoe, M. F. Land, and D. H. Ballard, "Eye guidance in natural vision: Reinterpreting salience," *J. Vis.*, vol. 11, no. 5, p. 5, May 2011.

[21] A. Borji and L. Itti, "State-of-the-Art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2013.

[22] N. D. B. Bruce, C. Wloka, N. Frosst, S. Rahman, and J. K. Tsotsos, "On computational modeling of visual saliency: Examining what's right, and what's left," *Vis. Res.*, vol. 116, pp. 95–112, Nov. 2015.

[23] Z. Bylinskii, E. M. DeGennaro, R. Rajalingham, H. Ruda, J. Zhang, and J. K. Tsotsos, "Towards the quantitative evaluation of visual attention models," *Vis. Res.*, vol. 116, pp. 258–268, Nov. 2015.

[24] A. C. Schütz, D. I. Braun, and K. R. Gegenfurtner, "Eye movements and perception: A selective review," *J. Vis.*, vol. 11, no. 5, p. 9, Sep. 2011.

[25] T. Foulsham and G. Underwood, "What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition," *J. Vis.*, vol. 8, no. 2, p. 6, Feb. 2008.

[26] W. Einhäuser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," *J. Vis.*, vol. 8, no. 14, p. 18, Nov. 2008.

[27] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[28] S. Chikkerur, T. Serre, C. Tan, and T. Poggio, "What and where: A Bayesian inference theory of attention," *Vis. Res.*, vol. 50, no. 22, pp. 2233–2247, Oct. 2010.

[29] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 20, 2008, pp. 241–248.

[30] A. Clavelli, D. Karatzas, J. Lladós, M. Ferraro, and G. Boccignone, "Modelling task-dependent eye guidance to objects in pictures," *Cognit. Comput.*, vol. 6, no. 3, pp. 558–584, Sep. 2014.

[31] A. Torralba, "Contextual priming for object detection," *Int. J. Comp. Vis.*, vol. 53, no. 2, pp. 153–167, 2003.

[32] Z. Bylinskii, A. Recasens, A. Borji, A. Oliva, A. Torralba, and F. Durand, "Where should saliency models look next?" in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 809–824.

[33] M. Kummerer, T. S. A. Wallis, L. A. Gatys, and M. Bethge, "Understanding low-and high-level contributions to fixation prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4789–4798.

[34] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 740–757, Mar. 2019.

[35] C. A. Rothkopf, D. H. Ballard, and M. M. Hayhoe, "Task and context determine where you look," *J. Vis.*, vol. 7, no. 14, p. 16, Jul. 2016.

[36] S. C.-H. Yang, D. M. Wolpert, and M. Lengyel, "Theoretical perspectives on active sensing," *Current Opinion Behav. Sci.*, vol. 11, pp. 100–108, Oct. 2016.

[37] B. B. Averbeck, "Theory of choice in bandit, information sampling and foraging tasks," *PLoS Comput. Biol.*, vol. 11, no. 3, Mar. 2015, Art. no. e1004164.

[38] L. Pessoa, "On the relationship between emotion and cognition," *Nature Rev. Neurosci.*, vol. 9, no. 2, pp. 148–158, 2008.

[39] K. C. Berridge and T. E. Robinson, "Parsing reward," *Trends Neurosci.*, vol. 26, no. 9, pp. 507–513, Sep. 2003.

[40] B. A. Anderson, "A value-driven mechanism of attentional selection," *J. Vis.*, vol. 13, no. 3, p. 7, Apr. 2013.

[41] E. Awh, A. V. Belopolsky, and J. Theeuwes, "Top-down versus bottom-up attentional control: A failed theoretical dichotomy," *Trends Cognit. Sci.*, vol. 16, no. 8, pp. 437–443, Aug. 2012.

[42] B. Tatler and B. Vincent, "Systematic tendencies in scene viewing," *J. Eye Movement Res.*, vol. 2, no. 2, pp. 1–18, Dec. 2008.

[43] B. W. Tatler and B. T. Vincent, "The prominence of behavioural biases in eye guidance," *Vis. Cognition*, vol. 17, nos. 6–7, pp. 1029–1054, Aug. 2009.

[44] M. Dorr, T. Martinetz, K. R. Gegenfurtner, and E. Barth, "Variability of eye movements when viewing dynamic natural scenes," *J. Vis.*, vol. 10, no. 10, p. 28, Aug. 2010.

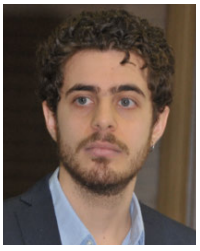
[45] Z. Hu, S. Li, C. Zhang, K. Yi, G. Wang, and D. Manocha, "DGaze: CNN-based gaze prediction in dynamic scenes," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 5, pp. 1902–1911, May 2020.

[46] J. M. Henderson and S. G. Luke, "Stable individual differences in saccadic eye movements during reading, pseudoreading, scene viewing, and scene search," *J. Experim. Psychol., Hum. Perception Perform.*, vol. 40, no. 4, p. 1390, 2014.
[47] G. Bargary, J. M. Bosten, P. T. Goodbourn, A. J. Lawrance-Owen, R. E. Hogg, and J. D. Mollon, "Individual differences in human eye movements: An oculomotor signature?" *Vis. Res.*, vol. 141, pp. 157–169, Dec. 2017.
[48] Y. Nagai, "From bottom-up visual attention to robot action learning," in *Proc. IEEE 8th Int. Conf. Develop. Learn.*, Jun. 2009, pp. 1–6.
[49] H. Rezazadegan Tavakoli, E. Rahtu, and J. Heikkilä, "Stochastic bottom–up fixation prediction and saccade generation," *Image Vis. Comput.*, vol. 31, no. 9, pp. 686–693, Sep. 2013.
[50] O. Le Meur and A. Coutrot, "Introducing context-dependent and spatially-variant viewing biases in saccadic models," *Vis. Res.*, vol. 121, pp. 72–84, Apr. 2016.
[51] T. D. Keech and L. Resca, "Eye movements in active visual search: A computable phenomenological model," *Attention, Perception, Psychophys.*, vol. 72, no. 2, pp. 285–307, Feb. 2010.
[52] U. Rutishauser and C. Koch, "Probabilistic modeling of eye movement data during conjunction search via feature-based attention," *J. Vis.*, vol. 7, no. 6, p. 5, Apr. 2007.
[53] D. Brockmann and T. Geisel, "The ecology of gaze shifts," *Neurocomputing*, vols. 32–33, no. 1, pp. 643–650, Jun. 2000.
[54] G. Boccignone and M. Ferraro, "Modelling gaze shift as a constrained random walk," *Phys. A, Stat. Mech. Appl.*, vol. 331, nos. 1–2, pp. 207–218, Jan. 2004.
[55] M. S. Cain, E. Vul, K. Clark, and S. R. Mitroff, "A Bayesian optimal foraging model of human visual search," *Psychol. Sci.*, vol. 23, no. 9, pp. 1047–1054, Sep. 2012.
[56] G. Boccignone and M. Ferraro, "Ecological sampling of gaze shifts," *IEEE Trans. Cybern.*, vol. 44, no. 2, pp. 266–279, Feb. 2014.
[57] P. Napoletano, G. Boccignone, and F. Tisato, "Attentive monitoring of multiple video streams driven by a Bayesian foraging strategy," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3266–3281, Nov. 2015.
[58] E. A. B. Over, I. T. C. Hooge, B. N. S. Vlaskamp, and C. J. Erkelens, "Coarse-to-fine eye movement strategy in visual search," *Vis. Res.*, vol. 47, no. 17, pp. 2272–2280, Aug. 2007.
[59] E. M. Kaya and M. Elhilali, "Modelling auditory attention," *Philos. Trans. Roy. Soc. B, Biol. Sci.*, vol. 372, no. 1714, 2017, Art. no. 20160101.
[60] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, p. 746, 1976.
[61] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *J. Acoust. Soc. Amer.*, vol. 26, no. 2, pp. 212–215, Mar. 1954.
[62] L. A. Ross, D. Saint-Amour, V. M. Leavitt, D. C. Javitt, and J. J. Foxe, "Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments," *Cerebral Cortex*, vol. 17, no. 5, pp. 1147–1153, Jun. 2006.
[63] G. A. Calvert, E. T. Bullmore, M. J. Brammer, R. Campbell, S. C. Williams, P. K. McGuire, P. W. Woodruff, S. D. Iversen, and A. S. David, "Activation of auditory cortex during silent lipreading," *Science*, vol. 276, no. 5312, pp. 593–596, Apr. 1997.
[64] K. V. Kriegstein, A. Kleinschmidt, P. Sterzer, and A.-L. Giraud, "Interaction of face and voice areas during speaker recognition," *J. Cognit. Neurosci.*, vol. 17, no. 3, pp. 367–376, Mar. 2005.
[65] E. Van der Burg, C. N. L. Olivers, A. W. Bronkhorst, and J. Theeuwes, "Audiovisual events capture attention: Evidence from temporal order judgments," *J. Vis.*, vol. 8, no. 5, p. 2, May 2008.
[66] H. M. Kondo, A. M. van Loon, J.-I. Kawahara, and B. C. J. Moore, "Auditory and visual scene analysis: An overview," *Philos. Trans. Roy. Soc. B, Biol. Sci.*, vol. 372, no. 1714, 2017.
[67] C. Kayser, C. I. Petkov, M. Lippert, and N. K. Logothetis, "Mechanisms for allocating auditory attention: An auditory saliency map," *Current Biol.*, vol. 15, no. 21, pp. 1943–1947, Nov. 2005.
[68] S. Onat, K. Libertus, and P. König, "Integrating audiovisual information for the control of overt attention," *J. Vis.*, vol. 7, no. 10, p. 11, Jul. 2007.
[69] G. Evangelopoulos, K. Rapantzikos, P. Maragos, Y. Avrithis, and A. Potamianos, "Audiovisual attention modeling and salient event detection," in *Multimodal Processing and Interaction*. Cham, Switzerland: Springer, 2008, pp. 1–21.
[70] A. Coutrot and N. Guyader, "An audiovisual attention model for natural conversation scenes," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 1100–1104.
[71] A. Coutrot and N. Guyader, "An efficient audiovisual saliency model to predict eye positions when looking at conversations," in *Proc. 23rd Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2015, pp. 1531–1535.
[72] A. Coutrot and N. Guyader, "Multimodal saliency models for videos," in *From Human Attention to Computational Attention*. Cham, Switzerland: Springer, 2016, pp. 291–304.
[73] G. Boccignone, V. Cuculo, A. D'Amelio, G. Grossi, and R. Lanzarotti, "Give ear to my face: Modelling multimodal attention to social interactions," in *Computer Vision—ECCV Workshops*, L. Leal-Taixé and S. Roth, Eds. Cham, Switzerland: Springer, 2019, pp. 331–345.
[74] H. R. Tavakoli, A. Borji, J. Kannala, and E. Rahtu, "Deep audio-visual saliency: Baseline model and data," in *Proc. ACM Symp. Eye Tracking Res. Appl.*, Jun. 2020, pp. 1–5.
[75] T. T. Hills, "Animal foraging and the evolution of goal-directed cognition," *Cognit. Sci.*, vol. 30, no. 1, pp. 3–41, Jan. 2006.
[76] J. Otero-Millan, S. L. Macknik, R. E. Langston, and S. Martinez-Conde, "An oculomotor continuum from exploration to fixation," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 15, pp. 6175–6180, Apr. 2013.
[77] C. A. Marlow, I. V. Viskontas, A. Matlin, C. Boydston, A. Boxer, and R. P. Taylor, "Temporal structure of human gaze dynamics is invariant during free viewing," *PLoS ONE*, vol. 10, no. 9, Sep. 2015, Art. no. e0139379.
[78] R. F. Green, "Bayesian birds: A simple example of Oaten's stochastic model of optimal foraging," *Theor. Population Biol.*, vol. 18, no. 2, pp. 244–256, Oct. 1980.
[79] J. McNamara, "Optimal patch use in a stochastic environment," *Theor. Population Biol.*, vol. 21, no. 2, pp. 269–288, Apr. 1982.
[80] J. McNamara and A. Houston, "A simple model of information use in the exploitation of patchily distributed food," *Animal Behav.*, vol. 33, no. 2, pp. 553–560, May 1985.
[81] K. A. Ehinger and J. M. Wolfe, "When is it time to move to the next map? Optimal foraging in guided visual search," *Attention, Perception, Psychophys.*, vol. 78, no. 7, pp. 2135–2151, Oct. 2016.
[82] S. Barthelmé, H. Trukenbrod, R. Engbert, and F. Wichmann, "Modeling fixation locations using spatial point processes," *J. Vis.*, vol. 13, no. 12, p. 1, Oct. 2013.
[83] P. Han, D. R. Saunders, R. L. Woods, and G. Luo, "Trajectory prediction of saccadic eye movements using a compressed exponential model," *J. Vis.*, vol. 13, no. 8, p. 27, 2013.
[84] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annu. Rev. Neurosci.*, vol. 18, no. 1, pp. 193–222, Mar. 1995.
[85] H. E. Egeth and S. Yantis, "VISUAL ATTENTION: Control, representation, and time course," *Annu. Rev. Psychol.*, vol. 48, no. 1, pp. 269–297, Feb. 1997.
[86] J. T. Serences and S. Yantis, "Selective visual attention and perceptual coherence," *Trends Cognit. Sci.*, vol. 10, no. 1, pp. 38–45, Jan. 2006.
[87] J. Fecteau and D. Munoz, "Salience, relevance, and firing: A priority map for target selection," *Trends Cognit. Sci.*, vol. 10, no. 8, pp. 382–390, Aug. 2006.
[88] P. C. Klink, P. Jentgens, and J. A. M. Lorteije, "Priority maps explain the roles of value, attention, and salience in goal-oriented behavior," *J. Neurosci.*, vol. 34, no. 42, pp. 13867–13869, Oct. 2014.
[89] L. Chelazzi, J. Eštočinová, R. Calletti, E. Lo Gerfo, I. Sani, C. D. Libera, and E. Santandrea, "Altering spatial priority maps via reward-based learning," *J. Neurosci.*, vol. 34, no. 25, pp. 8594–8604, Jun. 2014.
[90] E. Nelson, *Deep Audio-Visual Saliency: Baseline Model and Data*. Princeton, NJ, USA: Princeton Univ. Press, 1967.
[91] D. R. Brillinger, H. K. Preisler, A. A. Ager, J. G. Kie, and B. S. Stewart, "Employing stochastic differential equations to model wildlife motion," *Bull. Brazilian Math. Soc.*, vol. 33, no. 3, pp. 385–408, Nov. 2002.
[92] D. S. Lemons, *An Introduction to Stochastic Processes in Physics*. Baltimore, MD, USA: The Johns Hopkins Univ. Press, 2002.
[93] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, vol. 23. Berlin, Germany: Springer-Verlag, 2013.
[94] G. A. Breed, E. A. Golson, and M. T. Tinker, "Predicting animal home-range structure and transitions using a multistate Ornstein–Uhlenbeck biased random walk," *Ecology*, vol. 98, no. 1, pp. 32–47, Jan. 2017.
[95] K. J. Harris and P. G. Blackwell, "Flexible continuous-time modelling for heterogeneous animal movement," *Ecol. Model.*, vol. 255, pp. 29–37, Apr. 2013.
[96] R. H. MacArthur and E. R. Pianka, "On optimal use of a patchy environment," *Amer. Naturalist*, vol. 100, no. 916, pp. 603–609, Nov. 1966.
[97] E. L. Charnov, "Optimal foraging, the marginal value theorem," *Theor. Population Biol.*, vol. 9, no. 2, pp. 129–136, Apr. 1976.

[98] Y. Iwasa, M. Higashi, and N. Yamamura, "Prey distribution as a factor determining the choice of optimal foraging strategy," *Amer. Naturalist*, vol. 117, no. 5, pp. 710–723, May 1981.
[99] M. A. Rodríguez-Gironés and R. A. Vasquez, "Density-dependent patch exploitation and acquisition of environmental information," *Theor. Population Biol.*, vol. 52, no. 1, pp. 32–42, Aug. 1997.
[100] L. D. Kazimierski, G. Abramson, and M. N. Kuperman, "The movement of a forager: Strategies for the efficient use of resources," *Eur. Phys. J. B*, vol. 89, no. 10, p. 232, Oct. 2016.
[101] D. R. Cox and H. D. Miller, *The Theory Stochastic Processes*. Boca Raton, FL, USA: CRC Press, 2001.
[102] D. Insua, F. Ruggeri, and M. Wiper, *Bayesian Analysis of Stochastic Process Models*, vol. 978. Hoboken, NJ, USA: Wiley, 2012.
[103] M. Dorr, M. Bohme, T. Martinetz, K. Gegenfurtner, and E. Barth, "Variability of eye movements on natural videos," in *Proc. 8th Tubingen Perception Conf.*, Tubingen, Germany, Feb. 2005, p. 162.
[104] O. Le Meur and Z. Liu, "Saccadic model of eye movements for free-viewing condition," *Vis. Res.*, vol. 116, pp. 152–164, Nov. 2015.
[105] C. Xia, J. Han, F. Qi, and G. Shi, "Predicting human saccadic scanpaths based on iterative representation learning," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3502–3515, Jul. 2019.
[106] C. Xia and R. Quan, "Predicting saccadic eye movements in free viewing of Webpages," *IEEE Access*, vol. 8, pp. 15598–15610, 2020.
[107] W. Bao and Z. Chen, "Human scanpath prediction based on deep convolutional saccadic model," *Neurocomputing*, vol. 404, pp. 154–164, Sep. 2020.
[108] D. Zanca, S. Melacci, and M. Gori, "Gravitational laws of focus of attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jun. 4, 2019, doi: 10.1109/TPAMI.2019.2920636 .
[109] O. Le Meur and T. Baccino, "Methods for comparing scanpaths and saliency maps: Strengths and weaknesses," *Behav. Res. Methods*, vol. 45, no. 1, pp. 251–266, Mar. 2013.
[110] N. C. Anderson, F. Anderson, A. Kingstone, and W. F. Bischof, "A comparison of scanpath comparison methods," *Behav. Res. Methods*, vol. 47, no. 4, pp. 1377–1392, Dec. 2015.
[111] F. Cristino, S. Mathôt, J. Theeuwes, and I. D. Gilchrist, "ScanMatch: A novel method for comparing fixation sequences," *Behav. Res. Methods*, vol. 42, no. 3, pp. 692–700, Aug. 2010.
[112] H. Jarodzka, K. Holmqvist, and M. Nyström, "A vector-based, multidimensional scanpath similarity measure," in *Proc. Symp. Eye-Tracking Res. Appl. (ETRA)*, New York, NY, USA, 2010, pp. 211–218.
[113] R. Dewhurst, M. Nyström, H. Jarodzka, T. Foulsham, R. Johansson, and K. Holmqvist, "It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach," *Behav. Res. Methods*, vol. 44, no. 4, pp. 1079–1100, Dec. 2012.
[114] J. Pekkanen and O. Lappi, "A new and general approach to signal denoising and eye movement classification based on segmented linear regression," *Sci. Rep.*, vol. 7, no. 1, pp. 1–13, Dec. 2017.
[115] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. New York, NY, USA: Academic, 2013.
[116] N. Cliff, *Ordinal Methods for Behavioral Data Analysis*. London, U.K.: Psychology Press, 2014.
[117] M. R. Hess and J. D. Kromrey, "Robust confidence intervals for effect sizes: A comparative study of Cohen'sd and cliff's delta under non-normality and heterogeneous variances," in *Proc. Annu. Meeting Amer. Educ. Res. Assoc.*, 2004, pp. 1–30.
[118] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *J. Amer. Stat. Assoc.*, vol. 32, no. 200, pp. 675–701, Dec. 1937.
[119] P. B. Nemenyi, "Distribution-free multiple comparisons," Ph.D. dissertation, Dept. Math., Princeton Univ., Princeton, NJ, USA, 1963.
[120] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Jan. 2006.
[121] M. E. Wosniack, M. C. Santos, E. P. Raposo, G. M. Viswanathan, and M. G. E. da Luz, "The evolutionary origins of Lévy walk foraging," *PLoS Comput. Biol.*, vol. 13, no. 10, Oct. 2017, Art. no. e1005774.
[122] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Rev.-Neurosci.*, vol. 2, no. 3, pp. 1–11, 2001.
[123] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.
[124] M. Rubo and M. Gamer, "Social content and emotional valence modulate gaze fixations in dynamic scenes," *Sci. Rep.*, vol. 8, no. 1, pp. 1–11, Dec. 2018.
[125] T. V. Nguyen, Q. Zhao, and S. Yan, "Attentive systems: A survey," *Int. J. Comput. Vis.*, vol. 126, no. 1, pp. 86–110, Jan. 2018.
[126] J. S. Chung and A. Zisserman, "Out of time: Automated lip sync in the wild," in *Proc. Workshop Multi-View Lip-Reading (ACCV)*, 2016, pp. 251–263.
[127] J. S. Chung and A. Zisserman, "Lip reading in profile," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–11.
[128] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search," *Psychol. Rev.*, vol. 113, no. 4, p. 766, 2006.
[129] J. Schomaker, D. Walper, B. C. Wittmann, and W. Einhäuser, "Attention in natural scenes: Affective-motivational factors guide gaze independently of visual salience," *Vis. Res.*, vol. 133, pp. 161–175, Apr. 2017.
[130] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J. Vis.*, vol. 9, no. 12, pp. 1–27, 2009.
[131] P. Hu and D. Ramanan, "Finding tiny faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1522–1530.

GIUSEPPE BOCCIGNONE received the Laurea degree in theoretical physics from the University of Turin, Turin, Italy, in 1985. In 1986, he joined Olivetti Corporate Research, Ivrea, Italy. From 1990 to 1992, he has served as the Chief Researcher of the Computer Vision Laboratory, CRIAI, Naples, Italy. From 1992 to 1994, he held a Research Consultant position with the Research Labs, Bull HN, Milan, Italy, leading projects on biomedical imaging. In 1994, he joined as an Assistant Professor with the Dipartimento di Ingegneria dell'Informazione e Ingegneria Elettrica, University of Salerno, Fisciano, Italy. In 2008, he joined the Dipartimento di Informatica, University of Milan, Milan, where he is currently a Full Professor of statistics, natural interaction, and affective computing. His current research interests include visual attention, affective computing, Bayesian models, and stochastic processes for vision and the cognitive sciences.

VITTORIO CUCULO received the Ph.D. degree in mathematical sciences from the University of Milan, Milan, Italy, in 2017. Since 2017, he has been a Postdoctoral Researcher with the PHuSe Laboratory, Research Group, Department of Computer Science, University of Milan. His current research interests include affective computing, visual attention for health, positive technology, and signal processing.



ALESSANDRO D'AMELIO received the M.Sc. degree in computer science from the University of Milan, Milan, Italy, in 2017, where he is currently pursuing the Ph.D. degree in computer science. His current research interests include computational vision, affective computing, and Bayesian modeling.

GIULIANO GROSSI received the Ph.D. degree in computer science from the University of Milan, in 2000. He is currently an Assistant Professor in computer science with the University of Milan, where he has been an Assistant Professor with the Department of Computer Science, since 2001. As a member of the PHuSe Laboratory focused on affective and perceptive computing, his recent activities aim to apply both computer vision and machine learning techniques to human behavior understanding particularly refereed to social interaction, emotional state, and gaze analysis. He has authored 70 papers on international conferences and journals, and has been involved in several national and international projects concerning computer vision and Internet technology. His research interests also include sparse recovery in signal processing and dictionary learning with applications to face recognition and biosignal compression.

RAFFAELLA LANZAROTTI received the Ph.D. degree in computer science from the University of Milan, Milan, Italy, in 2003. Since 2004, she has been an Assistant Professor with the Department of Computer Science, University of Milan. Her current research interests include image and signal processing and affective computing, deepening issues concerning face images, such as face recognition and facial expression analysis, and physiological signal processing, such as ECG.