Give ear to my face: modelling multimodal attention to social interactions

Giuseppe Boccignone[0000−0002−5572−0924], Vittorio Cuculo[0000−0002−8479−9950], Alessandro D'Amelio[0000−0002−8210−4457], Giuliano Grossi[0000−0001−9274−4047], and Raffaella Lanzarotti[0000−0002−8534−4413]

PHuSe Lab, Department of Computer Science
Università degli Studi di Milano, Milano, Italy
{giuseppe.boccignone,vittorio.cuculo,alessandro.damelio}@unimi.it
{giuliano.grossi, raffaella.lanzarotti}@unimi.it

Abstract. We address the deployment of perceptual attention to social interactions as displayed in conversational clips, when relying on multimodal information (audio and video). A probabilistic modelling framework is proposed that goes beyond the classic saliency paradigm while integrating multiple information cues. Attentional allocation is determined not just by stimulus-driven selection but, importantly, by social value as modulating the selection history of relevant multimodal items. Thus, the construction of attentional priority is the result of a sampling procedure conditioned on the potential value dynamics of socially relevant objects emerging moment to moment within the scene. Preliminary experiments on a publicly available dataset are presented.

Keywords: Audio-visual attention · Social interaction · Multimodal perception.

1 Introduction

When humans are immersed in realistic, ecological situations that involve other humans, attention deployment strives for monitoring the behaviour, intentions and emotions of others even in the absence of a given external task [16]. Under such circumstances, the internal goal of the perceiver is to control attention so to maximise the implicit reward in focusing signals that bear social value [1].

Despite of experimental corroboration gained for such tendencies, their general modelling is far from evident (cfr., Section 2). Indeed, in order to put into work the mechanisms of selection, integration and sampling underlying the multifaceted phenomenon of attention, sensory systems have to master the flood of multimodal events (e.g., visual and audiovisual) captured in the external world. Thus, the research question we address in this note boils down to the following: is it possible to mine from behavioural data the implicit value of multimodal cues driving observer's motivation to spot socially interesting events in the scene?

Here, we propose a novel probabilistic model for grounding the inferential steps that lead to the prediction of a number of potential value-based attractors of multimodal attention (Section 3 and 4).

To this end, a clear case is represented by observers naturally viewing conversational videos conveying audiovisual information. Conversational clips are relatively controlled stimuli, while having the virtue of displaying real people embedded in a realistic dynamic situation [16]. Put simple, they allow affordable analysis and modelling of where and how people look when viewing such clips, namely, the fundamental questions entailed by spatiotemporal distribution of attention in a social context. Cogently, Foulsham *et al* [16] 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.

In what follows, to meet such experimental paradigm, we exploit the publicly available dataset by Coutrot and Guyader [13], who gathered data of eye-tracked subjects attending to conversational clips (Section 5). 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. As such it is suitable for both learning and testing the proposed model.

Model input, at the training stage, is represented by the audiovisual stream together with eye-tracking data. Inference is performed to obtain dynamic value-driven priority maps resulting from the competition of visual and audiovisual 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 model output is a number of attractors, namely clusters of potentially interest points sampled from priority maps, and suitable to guide attention control [29]. At the test stage, the latter can be compared with actual foci of attention selected by human subjects. Section 5 presents simulation results, and a conclusive discussion is given in Section 6.

2 Background and rationales

Whilst attentional mechanisms have been largely explored for vision systems, there is not much tradition as regards models of attention in the context of sound systems [18]. In vision, by and large, prominent models of attention foster a dichotomy between top-down and bottom-up control, with the former determined by current selection goals and the latter determined by physical salience [2, 31, 27]. Yet, the majority has retained a central place for low-level visual conspicuity [31, 5, 6], where the perceptual representation of the scene is usually epitomised in the form of a spatial saliency map, mostly derived bottom-up (early salience).

In a similar vein, such taxonomy has been assumed in the auditory attention field of inquiry. Since the seminal work by Kayser *et al* [19], efforts have been spent to model stimulus-driven attention to the auditory domain, by computing a visual saliency map of the spectrogram of an auditory stimulus (see [18] 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 [23, 15].

However, the top-down vs. bottom-up taxonomy of attentional control should be adopted with the uttermost caution (cfr., [2, 31]). On the one hand, the weakness of the bottom-up approach has been largely weighed up in the visual attention realm [31]). Early salience has only an indirect effect on attention by acting through recognised objects [14]. Thus, either object knowledge has been exploited (e.g., [9], in particular when dealing with faces [8], or contextual cues (e.g, the scene gist, [33]) for top-down tuning early salience. As a matter of fact, in the real world, most fixations are on task-relevant objects and this may or may not correlate with the saliency of regions. Further, the recent theoretical perspectives on active/attentive sensing promote a closed loop between an ideal observer, that extracts task-relevant information from a sequence of observations, and an ideal planner which specifies the actions that lead to the most informative observations [35]. The ultimate objective of active behaviour should be maximising the total rewards that can be obtained in the long term. On the other hand, there is a large body of evidence pointing at cases where strong selection biases cannot be explained by the physical salience of potential targets or by current selection goals. One such example is perceptual selection being biased towards objects associated with reward and selection history [2].

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 (independent of the internal state of the observer). However, the construct of "endogenous" attentional control is subtle since it conflates control signals that are "internal" (such as the motivation for paying attention to socially rewarding objects/events), "external" (induced by the given current task voluntarily pursued), and selection history (either learned or evolutionary inherited), which can prioritise items previously attended in a given context. If the ultimate objective of the attentive observer is total reward maximisation, one should clearly distinguish between "external" rewards (incentive motivation, e.g, monetary reward) and reward related to "internal" value. The latter has different psychological facets [3] including affect (implicit "liking" and conscious pleasure) and motivation (implicit incentive salience, "wanting"). Indeed, the selection of socially relevant stimuli by attention reflects the overall value of such selection [1].

3 Overview of the model

Under such circumstances, we generally assume attention as driven by goals $\mathcal{G}$ that, in turn, set the appropriate value $\mathcal{V}$ to events/objects occurring in the audiovisual scene. Also, in the work presented here, we assume that no explicit task is assigned to the perceiver; thus, value $\mathcal{V}$ is modulated by the "internal" goal (drive) towards spotting socially relevant objects/events. We consider two main inferential steps:

1. to infer a spatial-based priority map representation of the audio-visual landscape;

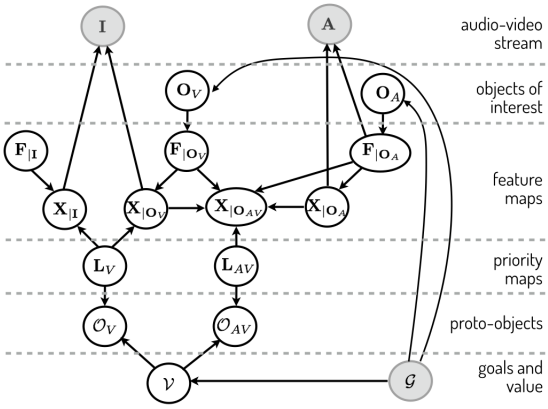

Fig. 1. An overall view of the model as a Probabilistic Graphical Model. Graph nodes denote RVs and directed arcs encode conditional dependencies between RVs. Grey-shaded nodes stand for RVs whose value is given. Time index t has been omitted for simplicity.

2. to exploit the priority map distributions, in order to sample value-based attractors suitable to guide attentional deployment.

Random variables (RVs), involved and their conditional dependencies are represented via the Probabilistic Graphical Model (PGM) outlined in Fig. 1.

Priority map representation. Perceptual spatial attention driven by multimodal cues mainly relies on visual and audio-visual priority maps, which we define as the RVs $\mathbf{L}_V$ and $\mathbf{L}_{AV}$, respectively. Formally, a priority map $\mathbf{L}$ is the matrix of binary RVs $l(\mathbf{r})$ 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. Thus, given the video and audio streams defining the audio-video landscape, $\{\mathbf{I}(t)\}$, $\{\mathbf{A}(t)\}$, respectively, a preliminary step is to evaluate at any time t, the posterior distributions $P(\mathbf{L}_V(t) \mid \mathbf{L}_V(t-1), \mathbf{I}(t))$ and $P(\mathbf{L}_{AV}(t) \mid \mathbf{L}_{AV}(t-1), \mathbf{A}(t), \mathbf{I}(t))$. The steps behind such estimate can be derived by resorting to the conditional dependencies defined in the PGM in Fig. 1. Backward inference $\{\mathbf{A}(t), \mathbf{I}(t)\} \rightarrow \{\mathbf{L}_V(t), \mathbf{L}_{AV}(t)\}$ stands upon a set of perceptual features $\mathbf{F}(t) = \{f(t)\}$ that can be estimated from the multimodal stream. From now on, for notational simplicity, we will omit time indexing t, unless needed.

As to the visual stream, we distinguish between two kinds of visual features: generic features $\mathbf{F}_{|\mathbf{I}}$ - such as edge, texture, colour, motion features-, and object-dependent features, $\mathbf{F}_{|\mathbf{O}}$. As to object-based features, these are to be learned by specifically taking into account the classes of objects that are likely to be relevant under the goal $\mathcal{G}$, via the distribution $P(\mathbf{O} \mid \mathcal{G})$. Here, where the task is free viewing/listening, and internal goals are biased towards social cues, the prominent visual objects are faces, $\mathbf{O}_V = \{face\}$. Both kinds of visual features, $\mathbf{F}_{|\mathbf{I}}$and $\mathbf{F}_{|\mathbf{O}}$, can be estimated in a feed-forward way. Note that in the literature

face information is usually referred to as a top-down cue [27] as opposed to bottom-up cues. However, much like physically driven features, they are phyletic features, and their distribution $P(\mathbf{F}_{|\mathbf{O}_V} \mid \mathbf{O}_V = \textit{face})$ is learnt by biological visual systems along evolution or in early development stages.

In order to be processed, features $\mathbf{F}_{|\mathbf{I}}$ and $\mathbf{F}_{|\mathbf{O}_V}$ need to be spatially organised in feature maps. A feature map $\mathbf{X}$ is a topographically organised map that encodes the joint occurrence of a specific feature at a spatial location [9]. It can be considered the probabilistic counterpart of a salience map [9] and it can be equivalently represented as a unique map encoding the presence of different object dependent features $\mathbf{F}_{f|\mathbf{O}_V}$, or a set of object-specific feature maps, i.e. $\mathbf{X} = \{\mathbf{X}_f\}$ (e.g., a face map, a body map, etc.). More precisely, $\mathbf{X}_f$ is a matrix of binary RVs $x(\mathbf{r})$ denoting whether feature f is present or not present at location $\mathbf{L} = \mathbf{r}$. Simply put, $\mathbf{X}_f$ is a map defining the spatial occurrence of $\mathbf{F}_{f|\mathbf{O}_V}$ or $\mathbf{F}_{f|\mathbf{I}}$. In our case, we need to estimate the posteriors $P(\mathbf{X}_{|\mathbf{I}} \mid \mathbf{F}_{|\mathbf{I}})$ and $P(\mathbf{X}_{|\mathbf{O}_V} \mid \mathbf{F}_{|\mathbf{O}_V})$.

As to the processing of audio, similarly to visual processing, auditory objects form across different analysis scales [29]. Formation of sound elements with contiguous spectro-temporal structure, relies primarily on local structures (e.g., onsets and offsets, harmonic structure, continuity of frequency over time), while social communication signals, such as speech, have a rich spectro-temporal structure supporting short-term object formation (e.g. formation of syllables). The latter are linked together over time through continuity and similarity of higher-order perceptual features, such as location, pitch, timbre and learned meaning. In our setting, the objects of interest $\mathbf{O}_A$ are represented by speakers' voices [16], 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). From a social perspective, 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 $\mathbf{r}$.

At this point, audio-visual perception has been cast in a spatial attention problem and priority maps $\mathbf{L}_V$and $\mathbf{L}_{AV}$ can be eventually estimated through distributions $P(\mathbf{L}_V(t) \mid \mathbf{L}_V(t-1), \mathbf{X}_{|\mathbf{I}}, \mathbf{X}_{|\mathbf{O}_V})$ and $P(\mathbf{L}_{AV}(t)|\mathbf{L}_{AV}(t-1), \mathbf{X}_{|\mathbf{O}_{AV}})$. Note that, in general, the representation entailed by a priority map differs from that provided at a lower level by feature maps $\mathbf{X}$ (or classic salience). 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 [9, 33]. Also, it can be designed to act as a form of short term memory to keep track of which potential targets have been attended. Thus, $\mathbf{L}(t)$ depends on both current perceptual inferences on feature maps at time t and priority at time $t-1$. Denote $\pi_{AV} = P(\mathbf{X}_{|\mathbf{O}_{AV}} \mid \mathbf{X}_{|\mathbf{O}_A}, \mathbf{X}_{|\mathbf{O}_V}, \mathbf{F}_{|\mathbf{O}_A}, \mathbf{F}_{|\mathbf{O}_V})$, $\pi_I = P(\mathbf{X}_{|\mathbf{I}} \mid \mathbf{F}_{|\mathbf{I}})$ and $\pi_{OV} = P(\mathbf{X}_{|\mathbf{O}_V} \mid \mathbf{F}_{|\mathbf{O}_V})$, $\pi_{L_V} = P(\mathbf{L}_V(t) \mid \mathbf{L}_V(t -$

$1), \mathbf{X}_{|\mathbf{I}}, \mathbf{X}_{|\mathbf{O}_V})$, $\pi_{L_{AV}} = P(\mathbf{L}_{AV}(t)|\mathbf{L}_{AV}(t-1), \mathbf{X}_{|\mathbf{O}_{AV}})$. Then,

$$\pi_{L_V}(t) \approx \alpha_V(\pi_I(t)\pi_{OV}(t)) + (1-\alpha_V)\pi_{L_V}(t-1), \tag{1}$$

$$\pi_{L_{AV}}(t) \approx \alpha_{AV}\pi_{AV}(t) + (1-\alpha_{AV})\pi_{L_{AV}}(t-1). \tag{2}$$

where α_V and α_{AV} weight the contribution of currently estimated feature maps with respect to previous priority maps.

Priority map dynamics requires an initial prior $P(\mathbf{L})$, which 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", [32, 20]), which can be modelled by assuming a Gaussian distribution located on the viewing center.

Sampling value-based attractors of multimodal attention. The next main inferential step involves the use of priority map distributions $\mathbf{L}^{(\ell)}$, ℓ being an index on $\{V, AV\}$, to sample attention attractors. Sampling is based on their value or potential reward $\mathcal{V}$ for the perceiver. In accordance with object-based attention approaches, we introduce proto-objects $\mathcal{O}_p^{(\ell)}$, where $p = 1, \cdots, N_P^{(\ell)}$, $N_P^{(\ell)}$ being the number of proto-objects detected in the priority map ℓ. These are the actual dynamic support for attention, conceived as the dynamic interface between attentive and pre-attentive processing [4]. Given a priority map $\mathbf{L}^{(\ell)}$, a set of proto-objects $\mathcal{O}^{(\ell)} = \{\mathcal{O}_p^{(\ell)}\}_{p=1}^{N_P^{(\ell)}}$ is computed. Each proto-object has a sparse representation in terms of a cluster of points $\{\mathbf{r}_{i,p}\}_{i=1}^{N_{I,p}^{(\ell)}}$ and parameters $\Theta_p = (\mathcal{M}_p^{(\ell)}, \theta_p^{(\ell)})$. In the general case, where the priority map distribution is a complex distribution with multiple modes (which is much likely to occur for $\mathbf{L}_V$) such parameters must be estimated. Here, the set $\mathcal{M}_p^{(\ell)} = \{m_p^{(\ell)}(\mathbf{r})\}_{\mathbf{r}\in\mathbf{L}^{(\ell)}}$ stands for a map of binary RVs indicating the presence or absence of proto-object p, and the overall map of proto-objects is given by $\mathcal{M}^{(\ell)} = \bigcup_{p=1}^{N_P^{(\ell)}} \mathcal{M}_p^{(\ell)}$, where $\mathcal{M}_p^{(\ell)} \bigcap \mathcal{M}_k^{(\ell)} = \emptyset, p \neq k$. Location and shape of the proto-object are parametrised via $\theta_p^{(\ell)}$. Assume independent proto-objects. In a first step we estimate the proto-object support map from the landscape, i.e., $\widehat{\mathcal{M}}^{(\ell)} \sim P(\mathcal{M}^{(\ell)}|\mathbf{L}^{(\ell)})$. Then, in a second step, $\widehat{\theta}_p^{(\ell)} \sim P(\theta_p^{(\ell)}(t)|\widehat{\mathcal{M}}_p^{(\ell)})$, location and shape parameters $\theta_p^{(\ell)} = (\mu_p^{(\ell)}, \Sigma_p^{(\ell)})$, $\mu_p^{(\ell)}$ are derived, $\Sigma_p^{(\ell)}$ being an elliptical representation of the proto-object support (location and axes).

As stated above, each proto-object relies on a sparse representation, i.e. the samples $\{\mathbf{r}_{i,p}\}_{i=1}^{N_{I,p}^{(\ell)}}$ representing candidate interest points (IPs). Sampling takes place conditionally on proto-object parameters $\theta_p^{(\ell)}$, and crucially it is modulated by value $\mathcal{V}^{(\ell)}$ that the perceiver is likely to gain from attending to the proto-object in map ℓ. Thus, considering $\mathcal{O}_p^{(\ell)}$ derived from the ℓ-th priority map,

$$\widetilde{\mathcal{O}}_p^{(\ell)} \triangleq \{w_{i,p}^{(\ell)}\mathbf{r}_{i,p}^{(\ell)}\}_{i=1}^{N_{I,p}^{(\ell)}} \sim P(\mathcal{O}_p^{(\ell)} \mid \theta_p^{(\ell)}, \mathcal{V}^{(\ell)}). \tag{3}$$

In Eq. 3, the posterior on $\mathcal{O}_p^{(\ell)}$ is a Gaussian distribution and the number of samples $N_p^{(\ell)}$ and the weight $w_{i,p}^{(\ell)}$ assigned to each particle $\mathbf{r}_{i,p}^{(\ell)}$ is a function of $\mathcal{V}^{(\ell)}$ attributed to the priority map ℓ.

Value $\mathcal{V}^{(\ell)}$, moment-to-moment updates according to the pdf $P(\mathcal{V}^{(\ell)}(t) \mid \mathcal{V}^{(\ell)}(t-1), \mathcal{G})$, depending on previous history and goal $\mathcal{G}$. Thus, by considering the time varying random vectors, $\mathcal{V}(t) = \{\mathcal{V}^{(\ell)}(t)\}$ (hidden continuous state) and $\mathcal{O}(t) = \{\mathcal{O}^{(\ell)}(t)\}$ (observable), value dynamics is best described by the following stochastic state-space system:

$$\widetilde{\mathcal{V}}(t) \sim P(\mathcal{V}(t) \mid \mathcal{V}(t-1), \mathcal{G}) \tag{4}$$

$$\widetilde{\mathcal{O}}(t) \sim P(\mathcal{O}(t) \mid \widetilde{\mathcal{V}}(t)) \tag{5}$$

Online inference is performed by solving the filtering problem $P(\mathcal{V}(t) \mid \mathcal{O}(1:t))$ under Markov assumption. This way current goal and selection history effects are both taken into account [2]. Such dynamics is set at the learning stage as detailed in the following section.

4 Current implementation of the model

The simulation of the model relies on a number of processing stages. At the lowest processing stages (in particular, face detection, audio-visual object detection), since we are dealing with feed-forward processes, thus we take advantage of efficient kernel-based methods and current deep neural network architectures. We give a brief sketch of the methods adopted.

Visual processing. In order to derive the physical stimulus feature map $\mathbf{X}_{|\mathbf{I}}$, we rely on the spatio-temporal saliency method proposed in [28] 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 [28], 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. There is a huge number of methods currently available: the one proposed by Hu and Ramanan [17] 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.

Audio visual processing. The features $\mathbf{F}_{|\mathbf{O}_A}$ used to encode the speech stream are the Mel-frequency cepstral coefficients (MFCC). The Mel-frequency cepstrum is highly effective in speech recognition and in modelling the subjective pitch and frequency content. 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 [10]; 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 [10] 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 [10, 11]), 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.

Priority maps and value-based proto-object sampling Priority maps are computed from feature maps, by simply using $\alpha = \alpha_V = \alpha_{AV}$, with $\alpha = 0.8$ experimentally determined via ROC analysis with respect to evaluation metrics (cfr. Section 5); such value grants higher weight to current information in order to account for changes in the audio-visual stream. From an experimental standpoint, we take into account four priority maps; namely, the visual priority map $\mathbf{L}_V$ as sensed from the video stream, the speaker/non-speaker maps, which we denote $\mathbf{L}_{AV_S}, \mathbf{L}_{AV_{NS}}$, and the one supporting the spatial prior $P(\mathbf{L}_V)$ (center bias), say $\mathbf{L}_{cb}$. To derive proto-objects from priority maps, markedly for estimating $\mathbf{L}_V(t)$, we need first to estimate their support $\mathcal{M}$ $\mathcal{M}(t) = \{m(\mathbf{r}, t)\}_{\mathbf{r}\in L}$, such that $m(\mathbf{r}, t) = 1$ if $P(\mathbf{L}(t)) > T_M$, and $m(\mathbf{r}, t) = 0$ 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. The procedure is based on the assumption that an informative proto-object is a relatively rare region and thus results in values which are in the tails of the distribution. Then, $\mathcal{M}(t) = \{\mathcal{M}_p(t)\}_{p=1}^{N_P}$ is obtained as $\mathcal{M}_p(t) = \{m_p(\mathbf{r}, t)|lab(B, \mathbf{r}, t) = p\}_{\mathbf{r}\in L}$, where the function *lab* labels $\mathcal{M}(t)$ around $\mathbf{r}$. We set the maximum number of proto-objects to $N_P = 15$, to retain the most important ones. The proto-object map provides the necessary spatial support for a 2D ellipse maximum-likelihood approximation of each proto-object, whose location and shape are parametrised as $\theta_p = (\mu_p, \Sigma_p)$ for $p = 1, \cdots, N_p$.

As previously stated, when sampling proto-object $\mathcal{O}_p^{(\ell)}$ (Eq. 3), with reference to the priority map $\mathbf{L}^{(\ell)}$, the number of samples $N_p^{(\ell)}$ and the weight $w_{i,p}^{(\ell)}$

assigned to each particle $\mathbf{r}_{i,p}^{(\ell)}$ is a function of value $\mathcal{V}^{(\ell)}$ attributed to the priority map ℓ. Thus, here the crucial issue is to determine the distribution update rule $P(\mathcal{V}^{(\ell)}(t) \mid \mathcal{V}^{(\ell)}(t-1), \mathcal{G})$. This is in general a difficult modelling issue, since value $\mathcal{V}(t) = \{\mathcal{V}^{(\ell)}(t)\}$ depends on current goals $\mathcal{G}$, either internal or external under a given task. However, the experimental eye-tracking data we use here are derived under a generic free-viewing task (external goal). Thus, we expect that attention allocation of observers in such context had been mainly driven by internal (endogenous) goals, most important the motivationally rewarding drive to deploy attention to conversational events/actors within the scene. Next, we exploit the *d*-separation property that holds for head-to-head dependencies in directed PGMs. When $\mathcal{O}^{(\ell)}$ and $\mathbf{L}^{(\ell)}$ are observed, it is possible to learn the dynamics of $\mathcal{V}(t)$ as the dynamics of a vector of time-varying parameters $\mathcal{V}^{(\ell)}(t)$. The latter control a function $g(\{\mathbf{L}^{(\ell)}\}, \mathcal{V}^{(\ell)}(t))$, which suitably combines the priority maps. This way the observation model in Eq. 5 can be expressed in the form $\widehat{g}(t) \sim P(g(\{\mathbf{L}^{(\ell)}\}, \mathcal{V}^{(\ell)}(t) \mid \widetilde{\mathcal{V}}(t))$. By assuming that $\widehat{g}(t)$ is an approximation of observers' attention allocation as summarised by the time-varying gaze heatmap $\mathcal{H}(t)$ computed from eye-tracked fixations, $\widehat{g}(t) \approx \mathcal{H}(t)$, then value dynamics can be learned by using $\mathcal{H}(t)$ as a ground-truth measurement. Generalising methods previously proposed for low-level saliency map weighting [25, 12, 30], value-state learning can be formulated as a time-varying regression, where the hidden-state evolution is that of the random vector of parameters $\mathcal{V}^{(\ell)}(t)$. To such end, at the learning stage we exploit the time-varying version of the Bayesian Lasso [24] to infer the joint hidden state dynamics $\prod_\ell P(\mathcal{V}^{(\ell)}(t) \mid \mathcal{V}^{(\ell)}(t-1))$ - under the assumption of independence between Gaussian distributed parameters - by using $\mathcal{H}(t)$ obtained from a subset of observers.

For the simulations and results presented in the following, the number of points $N^{(\ell)}$ to be sampled from priority map ℓ, is set as $N^{(\ell)} = \overline{\mathcal{V}}^{(\ell)}(t) N_{tot}$, where $N_{tot} = 500$ is the total number of points to be sampled and $\overline{\mathcal{V}}^{(\ell)}(t) = E\left[\mathcal{V}^{(\ell)}(t) \mid \mathcal{V}^{(\ell)}(t-1)\right]$ the value conditional expectation on map ℓ. Analogously, the weight $w_{i,p}^{(\ell)}$ assigned to each particle $\mathbf{r}_{i,p}^{(\ell)}$, is determined as $w_{i,p}^{(\ell)} = \overline{\mathcal{V}}^{(\ell)}(t) P(\mathbf{r}_{i,p}^{(\ell)})$ (cfr., Eq. 3). One typical result that shows the overall process at a glance is outlined in Fig. 4. The effect of value-based sampling is evident in the number of sampled points for the different priority maps; along a conversation, as expected, audio-visual events captured by $\mathbf{L}_{AV}$ are granted higher value with respect to the visual priority map $\mathbf{L}_V$ and the initial center bias prior $P(\mathbf{L})$. Figure 5 shows a snapshot of model output, where value assigned to sampled points is explicitly shown (colour coded).

5 Simulations

Stimuli and eye-tracking data. The adopted dataset [13] consists of 15 one-shot conversation scenes from French movies, involving two to four different actors for each scene. The duration of the videos goes from 12 to 30 seconds, with a resolution of 720×576 pixels at a frame rate of 25 fps. The dataset includes eye-tracking recordings in four different auditory conditions, but for the purposes of

[t]1

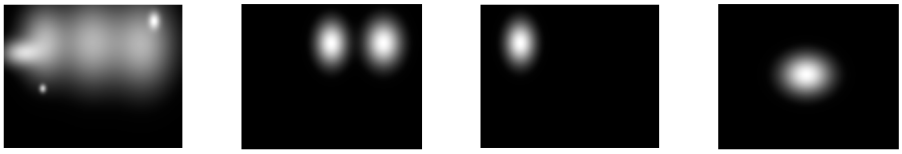

Fig. 2.

[t]

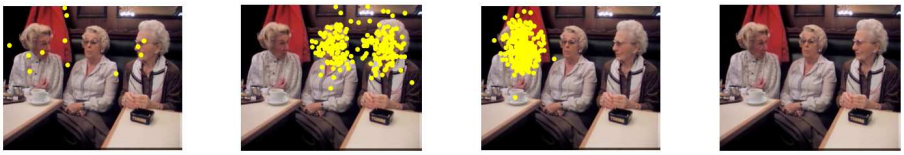

Fig. 3.

Fig. 4. Probability density functions (a) and value-based sampling (b) related to priority maps (top to bottom) $\mathbf{L}_V$ (visual), $\mathbf{L}_{AV}$ (non-speaker), $\mathbf{L}_{AV}$ (speaker); for convenience, the initial prior $P(\mathbf{L})$ (center bias, bottom panel) is also shown.

our work, the one with the original audio information has been employed. The experiment involved 18 different participants, all French native speakers and not aware of the purpose of the experiment. The eye-tracker system recorded eye positions at 1000 Hz, downsampled in accordance with the video frame rate, with a median of 40 raw consecutive eye positions [13].

Evaluation. In the present study 13 of the 15 clips were used; the *jeuxinterdits* video was excluded due to anomalous false positives rising in the face detection step (which is not matter of assessment here); the *fetecommence* clip was used for preliminary validation of parameter tuning and it was not included in the final performance assessment reported below. For each clip, the 18 observers have been randomly assigned to training (11) and test (7) sets. The learning stage, as previously discussed, was devoted to learn, from observers in the training set and for each video, the hidden state dynamics of value parameters $\{\mathcal{V}^{(\ell)}(t)\}$, $\prod_\ell P(\{\mathcal{V}^{(\ell)}(t)\} \mid \{\mathcal{V}^{(\ell)}(t-1)\})$, governing the time-varying Bayesian Lasso (Section 4). The eye-tracked data in the training set were used as the targets for supervised learning. To such end, the time-varying empirical distribution of observers' fixations was derived (the ground truth); the model-based distribution (apt to predict attention deployment and to be matched against the empirical one) was blindly obtained via standard Kernel Density Estimation on sampled points (cfr., Fig. 5). In the test stage, the empirical and model-based distributions have been eventually compared by adopting four widely adopted standard evaluation metrics [7]: Area under ROC Curve (AUC), Pearson's Correlation Coefficient (CC), Normalized Scanpath Saliency (NSS) and Kullback-Leibler divergence (KL). AUC is the most commonly-used metric for this kind of evaluation,

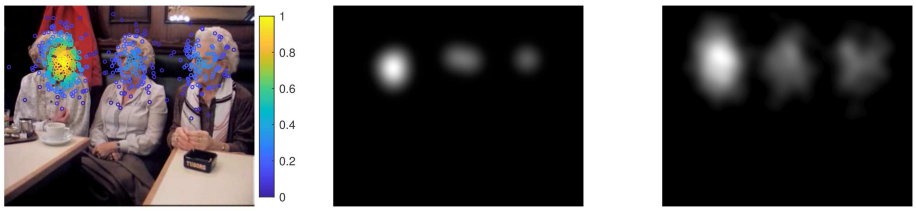

Fig. 5. Comparing human vs. model. From left to right: the overall sampling result overlaid on the video frame, colour representing value assigned to sampled points; the empirical distribution of human fixations; the model kernel-based distribution of sampled points.

mainly driven by high-valued predictions and largely ambivalent of low-valued false positives. CC is a linear correlation between the prediction and ground truth distributions, and treats false positives and false negatives symmetrically. NSS is discrete approximation of CC that is additionally parameter-free and operates on raw fixation locations. KL has a natural interpretation where goal is to approximate a target distribution while highly penalising mis-detections. The theoretical best performance limits of such metrics are 0.92, 1.00, 3.29 and 0 for AUC, CC, NSS and KL, respectively [7]. The overall quantitative evaluation of the simulation is summarised in Table 1, in terms of the mean value of the four metrics for each video over 7 model simulation trials.

Table 1. Mean value (and standard dev.) of the metric scores obtained for each video on the test set.

Video	**KLD**	**NSS**	**AUC**	**CC**
arrogants	1.25 ± 0.78	2.61 ± 0.54	0.90 ± 0.05	0.63 ± 0.12
conversation	1.80 ± 1.89	2.53 ± 0.66	0.85 ± 0.09	0.65 ± 0.19
equipier	1.22 ± 0.70	2.25 ± 0.69	0.83 ± 0.08	0.61 ± 0.17
hommedeneuve	1.07 ± 0.70	2.58 ± 0.53	0.87 ± 0.07	0.66 ± 0.12
hommeface	1.39 ± 1.27	2.69 ± 0.80	0.81 ± 0.09	0.70 ± 0.22
jeuxdenfant	1.04 ± 1.16	2.94 ± 0.36	0.84 ± 0.06	0.80 ± 0.09
moustacheassis	1.73 ± 1.29	2.58 ± 0.53	0.89 ± 0.06	0.63 ± 0.12
moustachepolicier	1.38 ± 0.79	2.32 ± 0.70	0.87 ± 0.09	0.57 ± 0.15
periljeune	1.58 ± 1.19	2.15 ± 0.57	0.87 ± 0.06	0.50 ± 0.18
pleincoeurbistrot	2.15 ± 1.86	2.53 ± 0.66	0.83 ± 0.08	0.66 ± 0.18
quatrevingtdixneuf	1.36 ± 0.19	1.99 ± 0.51	0.90 ± 0.04	0.50 ± 0.10
saveurspalais	1.36 ± 1.01	2.63 ± 0.49	0.88 ± 0.06	0.66 ± 0.10
unsoir	1.85 ± 1.79	2.39 ± 0.65	0.86 ± 0.10	0.58 ± 0.17
MEAN	1.48 ± 1.23	2.48 ± 0.6	0.86 ± 0.08	0.63 ± 0.15

Figure 6 provides an interesting snapshot of the evolution over time (video frames) of the four metrics. The transition to the onset of a speech event results

in higher uncertainty among observers (and thus model prediction) is captured by the absolute minimum of AUC, CC, NSS and the maximum KL; uncertainty is reduced after such onset, and scores evolve toward a steady-state of valuable performance with respect to their theoretical limits. Beyond notable results achieved in simulations, it is worth remarking that the comparison condition we adopted is somehow unfair with respect to the model. The ground truth is derived from actual observers' fixations, whilst model-based distribution is computed from the sampled points that only represent candidate fixation points, conceived to be subsequently exploited for deciding the actual gaze shift (which also explains in Fig. 5 the slightly bigger spread of model distributions with respect to the ground truth ones) .

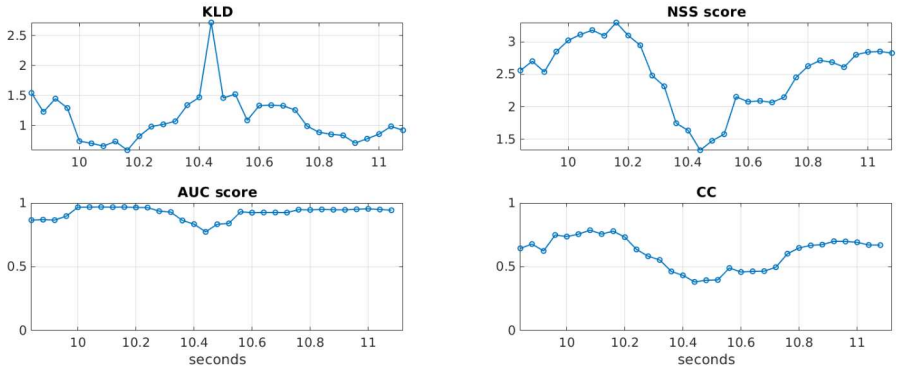

Fig. 6. A snapshot of metrics evolution after the $10th$ second of the video *"Faces conversation"*, when a change of speaker occurs.

6 Conclusions

This study gauged the importance of social information on attention in terms of gaze allocation when perceiving naturalistically rich, multimodal dynamic scenes. Hitherto the problem of modelling the behaviour of active observers in such context has seldom been taken into consideration, in spite of the exponentially growing body of audio-visual data conveying social behaviour content. The involvement of value is still in its infancy in the attention field of inquiry [31], as opposed to salience and objects that have been largely addressed both in psychology [27] and computer vision [5].

Preliminary experimental results show that the model is suitable to infer from behavioural data the implicit value of audio-visual cues driving observer's motivation to spot socially interesting events in the scene. The model is conceived in a probabilistic framework for object-based multimodal attention control [29].

Such formulation is far from evident. The notion that observers' visual attention is driven towards potential objects in the scene has been widely exploited, whilst sound might not always be allocated between objects; it could be conductive to multiple objects or to no object. Yet, the social context provides a thorough understanding of what is a potential auditory object and promotes the segmentation of ecologically consistent and valuable audio-visual entities (e.g., a speaking person). A mean to ground consistency has been synchronisation between audio and visual events an issue that has been previously addressed, e.g., [26, 12, 21]. To such end we have adapted to our framework recent results gained by deep network techniques [10, 11]. As a result, spatially-based probabilistic priority maps are built-up from the visual and auditory objects across different analysis scales. These maps are dynamic loci that moment-to-moment compete on the base of their activities. The latter are formalised in terms of value-driven proto-object sampling, to generate attractors for attention deployment. Cogently, sampling is conditional on the value dynamics (current history and "internal" goals) of the perceiver. This choice is consistent with theoretical model building trends [35] positing active attentional deployment as the problem of maximising the total rewards that can be gained by the active perceiver. Further, the broader perspective of "internal" value/reward, as brought forward by socially relevant stimuli [1, 3], paves the way to a wider dimension of attentional processing, e.g. including affective modulation of attention.

Value attribution dynamics is learnt on a video clip on the basis of eye-tracked gaze allocation of a number of observer and can be used, at the testing stage, to predict attentional deployment of novel observers on the same clip. In this respect, one may raise the issue that the inferred viewing behaviour might not generalise to novel kind of stimuli content. This objection is true but with reference to the specific unfolding in time of value dynamics on a given video clip, as provided by the current regression-like implementation of Eqs 4 and 5. Even so, on the one hand, the model simulation as such (that is, in the same experimental setting we have presented here) could be applied to a variety of investigations of social attention behaviour in groups that are likely to differentiate with respect to the given stimulus (e.g., clinical populations). On the other hand, the model captures and, cogently, quantifies across the different video clips some general patterns of value attribution in attentional allocation: low-level, physically driven cues (early salience) play a marginal role when social cues are present; effects due to spatial tendencies, such as the center bias, are relevant at the onset of the stimulus, and rapidly decrease in time; attention deployment is rapidly devoted to people in the video, speakers bearing the highest value for the observer. Yet, issues of generalisation across different video clips were out of the scope of the study presented here and are currently part of ongoing research.

A key feature of our approach is the inherent stochasticity of the model (sampling). This is *per se* apt to account for either observers' inter- and intra-variability in audio-visual perceptual tasks. More generally, randomness in actual attention deployment (as eventually gauged through gaze shifts) is likely to be originated from endogenous stochastic variations that affect each stage between

a perceptual event and the motor response: sensing, information processing, movement planning and executing [32]. At bottom, it should be always kept in mind that the actual process involving eye movement behaviour is a closed loop between an ideal observer, that extracts task-relevant information from a sequence of observations, and an ideal planner which specifies the actions that lead to the most rewarding sampling [35]. The latter issue involving the actual gaze shift (motor action) is often neglected in the literature [31]. In point of fact, oculomotor behaviour encapsulates either noisy motor responses and systematic tendencies in the manner in which we explore scenes with our eyes [32], albeit modulated by the semantic category of the stimulus [20]. Overall and most important, this approach paves the way to the possibility of treating reward-based visual exploration strategies in the framework of stochastic *information foraging* [34, 4, 22], a promising research line for which the model presented here is likely to offer a sound basis.

References

1. Anderson, B.A.: A value-driven mechanism of attentional selection. Journal of vision **13**(3) (2013)
2. Awh, E., Belopolsky, A.V., Theeuwes, J.: Top-down versus bottom-up attentional control: A failed theoretical dichotomy. Trends in cognitive sciences **16**(8), 437–443 (2012)
3. Berridge, K.C., Robinson, T.E.: Parsing reward. Trends in neurosciences **26**(9), 507–513 (2003)
4. Boccignone, G., Ferraro, M.: Ecological sampling of gaze shifts. IEEE Trans. on Cybernetics **44**(2), 266–279 (Feb 2014)
5. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(1), 185–207 (2013)
6. Bruce, N.D., Wloka, C., Frosst, N., Rahman, S., Tsotsos, J.K.: On computational modeling of visual saliency: Examining what's right, and what's left. Vision research **116**, 95–112 (2015)
7. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? IEEE Trans. on Pattern Analysis and Machine Intelligence pp. 1–1 (March 2018)
8. Cerf, M., Harel, J., Einhäuser, W., Koch, C.: Predicting human gaze using low-level saliency combined with face detection. Advances in neural information processing systems **20** (2008)
9. Chikkerur, S., Serre, T., Tan, C., Poggio, T.: What and where: A bayesian inference theory of attention. Vision research **50**(22), 2233–2247 (2010)
10. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: Workshop on Multi-view Lip-reading, ACCV (2016)
11. Chung, J.S., Zisserman, A.: Lip reading in profile. BMVC (2017)
12. Coutrot, A., Guyader, N.: An efficient audiovisual saliency model to predict eye positions when looking at conversations. In: 23rd European Signal Processing Conference. pp. 1531–1535 (Aug 2015)
13. Coutrot, A., Guyader, N.: How saliency, faces, and sound influence gaze in dynamic social scenes. Journal of vision **14**(8), 5–5 (2014)

14. Einhäuser, W., Spain, M., Perona, P.: Objects predict fixations better than early saliency. Journal of Vision **8**(14) (2008). https://doi.org/10.1167/8.14.18, http://www.journalofvision.org/content/8/14/18.abstract
15. Evangelopoulos, G., Rapantzikos, K., Maragos, P., Avrithis, Y., Potamianos, A.: Audiovisual attention modeling and salient event detection. In: Multimodal Processing and Interaction, pp. 1–21. Springer (2008)
16. Foulsham, T., Cheng, J.T., Tracy, J.L., Henrich, J., Kingstone, A.: Gaze allocation in a dynamic situation: Effects of social status and speaking. Cognition **117**(3), 319–331 (2010)
17. Hu, P., Ramanan, D.: Finding tiny faces. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1522–1530. IEEE (2017)
18. Kaya, E.M., Elhilali, M.: Modelling auditory attention. Phil. Trans. R. Soc. B **372**(1714), 20160101 (2017)
19. Kayser, C., Petkov, C.I., Lippert, M., Logothetis, N.K.: Mechanisms for allocating auditory attention: an auditory saliency map. Current Biology **15**(21), 1943–1947 (2005)
20. Le Meur, O., Coutrot, A.: Introducing context-dependent and spatially-variant viewing biases in saccadic models. Vision Research **121**, 72–84 (2016)
21. Nakajima, J., Sugimoto, A., Kawamoto, K.: Incorporating audio signals into constructing a visual saliency map. In: Pacific-Rim Symposium on Image and Video Technology. pp. 468–480. Springer (2013)
22. Napoletano, P., Boccignone, G., Tisato, F.: Attentive monitoring of multiple video streams driven by a bayesian foraging strategy. IEEE Trans. on Image Processing **24**(11), 3266 – 3281 (Nov 2015)
23. Onat, S., Libertus, K., König, P.: Integrating audiovisual information for the control of overt attention. Journal of Vision **7**(10), 11–11 (2007)
24. Park, T., Casella, G.: The bayesian lasso. Journal of the American Statistical Association **103**(482), 681–686 (2008)
25. Rahman, I.M., Hollitt, C., Zhang, M.: Feature map quality score estimation through regression. IEEE Trans. on Image Processing **27**(4), 1793–1808 (2018)
26. Rodríguez-Hidalgo, A., Peláez-Moreno, C., Gallardo-Antolín, A.: Towards multimodal saliency detection: An enhancement of audio-visual correlation estimation. In: Proc. 16th Int. Conf. on Cognitive Informatics & Cognitive Computing. pp. 438–443. IEEE (2017)
27. Schütz, A., Braun, D., Gegenfurtner, K.: Eye movements and perception: A selective review. Journal of Vision **11**(5) (2011)
28. Seo, H., Milanfar, P.: Static and space-time visual saliency detection by self-resemblance. Journal of Vision **9**(12), 1–27 (2009)
29. Shinn-Cunningham, B.G.: Object-based auditory and visual attention. Trends in cognitive sciences **12**(5), 182–186 (2008)
30. Suda, Y., Kitazawa, S.: A model of face selection in viewing video stories. Scientific Reports **5**, 7666 (2015)
31. Tatler, B., Hayhoe, M., Land, M., Ballard, D.: Eye guidance in natural vision: Reinterpreting salience. Journal of vision **11**(5) (2011)
32. Tatler, B., Vincent, B.: The prominence of behavioural biases in eye guidance. Visual Cognition **17**(6-7), 1029–1054 (2009)
33. Torralba, A.: Contextual priming for object detection. Int. J. of Comp. Vis. **53**, 153–167 (2003)
34. Wolfe, J.M.: When is it time to move to the next raspberry bush? Foraging rules in human visual search. Journal of vision **13**(3), 10 (2013)

35. Yang, S.C.H., Wolpert, D.M., Lengyel, M.: Theoretical perspectives on active sensing. Current Opinion in Behavioral Sciences **11**, 100–108 (2016)