2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Congress Center Hamburg
Sept 28 - Oct 2, 2015. Hamburg, Germany

Designing an Artificial Attention System for Social Robots

Pablo Lanillos[a], João Filipe Ferreira[a] and Jorge Dias[a,b]

Abstract— **In this paper, we introduce the main components comprising the action-perception loop of an overarching framework implementing artificial attention, designed to fulfil the requirements of social interaction (i.e., reciprocity, and awareness), with strong inspiration on current theories in functional neuroscience. We demonstrate the potential of our framework, by showing how it exhibits coherent behaviour without any inbuilt prior expectations regarding the experimental scenario. Current research in cognitive systems for social robots has suggested that automatic attention mechanisms are essential to social interaction. In fact, we hypothesise that enabling artificial cognitive systems with middleware implementing these mechanisms will empower robots to perform adaptively and with a higher degree of autonomy in complex and social environments. However, this type of assumption is yet to be convincingly and systematically put to the test. The ultimate goal will be to test our working hypothesis and the role of attention in adaptive, social robotics.**

I. Introduction

In the past decades, robotics has drawn a substantial deal of inspiration from neuroscience and psychology in the attempt to properly address the action-perception loop, in particular with "theory of mind" and evolutionary and developmental approaches [1], [2], which in turn have brought attention to the limelight. The underlying rationale is as follows: by developing attentional systems with some of the functionalities found in the human brain, robots will not only be able to exhibit behaviours that resemble those of their interlocutors, but also gain additional advantages such as being able to respond adaptively to the environment [3]. This is important in order to be able to launch the foundations of processes such as empathy, mirroring and reciprocity, given that the human interlocutor will most certainly build his/her own mirrored representation of the robot actions and intentions [4], [5], [6]. Consequently, recent research lines have suggested that automatic attentional mechanisms are a fundamental foundation for implementing robotic intelligence in the development of social robots [3], [6], [5]. As opposed to tailor-made solutions mostly focussed on solving very specific cognitive tasks, lacking the traits of adaptive behaviour that would allow robots to function in open-ended scenarios, we advocate an approach for attention system design that incorporates as much of what is known of attentional processes in the brain as possible to adaptively deal with uncertain scenarios.

The primary purpose of this paper is to provide a general unified design of the robotic attentional mechanism by bringing together various elements of previous works in neuropsychology and robotics theories and applications. Furthermore, we provide details on some of its most important modules, and discuss overall functionality. Computational modelling has been tackled by resorting to probabilistic techniques such as hierarchical Bayesian programming [7] and probabilistic state machines [8]. The probabilistic framework allows the robot to deal with the uncertainty inherent to the action-perception loop, fundamentally relating to recent studies about how the human brain deals with these processes [9], additionally providing an implicit methodology for signal fusion and modulation, and also for adaptive interaction. Moreover, the proposed framework assumes attention as a multisensory process - it is currently designed for visuoauditory perception, but is intended to be generalisable to other important senses, such as touch or olfaction.

The remainder of this paper is structured as follows. Section II describes the main motivations of this work, analysing key theories from neuroscience. Section III presents related work already available in current robots and artificial cognitive systems. Section IV proposes an architecture for attention, provides details of some of its most important components and their mathematical foundation. Section V analyse by simulation and experimentation the current implementation of the attentional system. Finally, section VI discusses the potential benefits of using the proposed design and possible alternatives and improvements.

II. Bioinspired Foundations

When analysing human cognitive impairments or disorders, attention appears to be one of the most important skills to achieve correct social interaction [10], because it enables activities such as learning, visual search, non-verbal and verbal interaction, and is also one of the key processes underlying intentional inference. Currently, however, cognitive systems in robots have not yet tackled this problem comprehensively and generally enough [3]. Consequently, in terms of attention, robots should simultaneously be capable of:

1) behaving in a socially reciprocal fashion, by attending to important social cues as a human would when directly interacting within his/her social space, and by maintaining sequences of attentional behaviours regulating basic interaction activities such as joint attention;

This work was supported by the Portuguese Foundation for Science and Technology (FCT) and by the European Commission via the COMPETE programme [project grant number FCOMP-01-0124-FEDER-028914, FCT Ref. PTDC/EEI-AUT/3010/2012].

[a] AP4ISR team, Institute of Systems and Robotics (ISR) Dept. of Electrical & Computer Eng., University of Coimbra. Pinhal de Marrocos, Pólo II, 3030-290 COIMBRA, Portugal jfilipe@isr.uc.pt.

[b] Jorge Dias is also with Khalifa University of Science, Technology, and Research Abu Dhabi 127788, UAE.

2) attending to unexpected stimuli that will help maintain a high degree of adaptability and responsiveness to changes in the current context that bear behavioural relevance.

Attention is the process whereby an agent allocates perceptual resources to analyse a subset of the surrounding world in detriment of others [3]. It is, therefore, a strategic and rather complex data-handling process that allows the processing of an unmanageable amount of information (sensory and otherwise) to become tractable. Several key theories from neuroscience, in favour of which a considerable amount of evidence has been amassed, have served as the main motivations for the proposed framework:

- Neurophysiologists have identified two highly interconnected attentional processes in the human brain: (1) a top-down (i.e., goal-oriented) modulation of bottom-up (i.e., stimulus-driven - e.g., saliency) attentional capture by targets versus distractors that is believed to be implemented by what has been called the *dorsal attention system*. [11], [12], [13]; and (2) a coordinated attentional process consisting of bottom-up attentional capture by behaviourally relevant distractors (e.g., unexpected stimuli) that is believed to be implemented by the *ventral attention system* of the human brain, and is filtered by behavioural valences to reorient attention by resetting the current attentional set accordingly [12], [13].
- Graziano et al. [14], [15] have proposed the "awareness theory", in which the brain is suggested to possess functional sites devoted to building a simplified, schematic model of the current state of the complex data-handling process of attention, which would serve as a model of awareness. Awareness would therefore allow the brain to understand attention, its dynamics, and its consequences. Moreover, they posit that more than a single schematic model of this sort, which they named the "attention schema", may be built using the same "machinery", namely for attributing an attentional state to oneself or to others. These simplified representations can be used to infer and predict intention and goals for both the self and others, serving as a support to cognitive processes such as those described by theories such as the "theory of mind". The "attention schema machinery" would build its simplified representation of an attentional state of an agent by using (accessed or inferred) knowledge of cues such as gaze direction, facial expression, body language, prior knowledge on the agent, location of salient objects, etc.
- Joint attention (JA) is a primal non-verbal interactive and cyclic process established between humans [3], which we believe is an integral part of the attentional process as a whole, by attributing to it its social trait. The JA interaction can be described, using an example, as follows: while playing with his father, a child stares at his parent when shown a toy ("initiate joint attention" – IJA – by the father), then will gaze a toy ("respond to joint attention" – RJA – by the child) and then again to his father in order to acknowledge that the other has understood that both are "talking about the same object" ("acknowledge joint attention" – AJA);
- Spivey, Richardson and Fitneva [16] stated that eye fixations serve as cognitive links, in the form of lists of deictic pointers bonded to spatial indices, between internal and external objects and events, suggesting that attention is used for organising relatively high-level cognitive processes. We propose that these lists could take an integral part in organising the set salient objects processed by the "attention schema" of Graziano et al.

III. Related Work in Robotics

Automatic attention hides multiple challenges that have been approached from different points of view. The most common approach is to basically model only the stimulus-driven, bottom-up aspect of attention using a saliency map that codifies the relevance of each location or entity based on the local contrast of low-level features [17], [5], [18], and then making this model compete with other goal-directed behaviours modelled separately [3]. Another approach, still focussing on bottom-up attention, is information theoretic modelling, where entropic or surprise measures provide the most probable locations [19], [20]. However, as mentioned in section II, attention is also known to be modulated by goal-directed signals, which has spurred new research efforts attempting to tackle this issue [21], [22], [23]. Attentional goals are also known to be informed by the environmental context, leading to research such as [24], and also by the object of interest for a specific task, leading to solutions that include modulation of attention via feedback through object segmentation and tracking [22], thereby closing the action-perception loop. Additionally, overt attention (i.e. active perception) is still a challenging task in terms of design and quantitative evaluation, due to its scene-dependent nature [25], [26].

On the other hand, defining the cognitive architecture or the computational model of a robot for general-purpose HRI is a difficult task, although developmental robotics give us the methodology to build cognitive abilities incrementally. Instead of defining specific solutions for each task, current research has favoured holistic solutions that build on sets of atomic functionalities [27]. The "theory of mind" applied to robots [2] opened the window for multiple biologically-inspired cognitive models. Surveys such as [28], [1] describe the latest approaches in cognitive developmental robotics. The role played by attention architectures in these holistic approaches, however, while having been *assumed* to be essential (as seen in the plethora of attention-related research in robotics summarised above), has yet to be convincingly and systematically demonstrated as such [3], [6].

IV. Attentional Architecture

Figure 1 shows the overall framework for the proposed system. There are four overarching interconnected modules, which will be detailed in the following subsections: (1) the

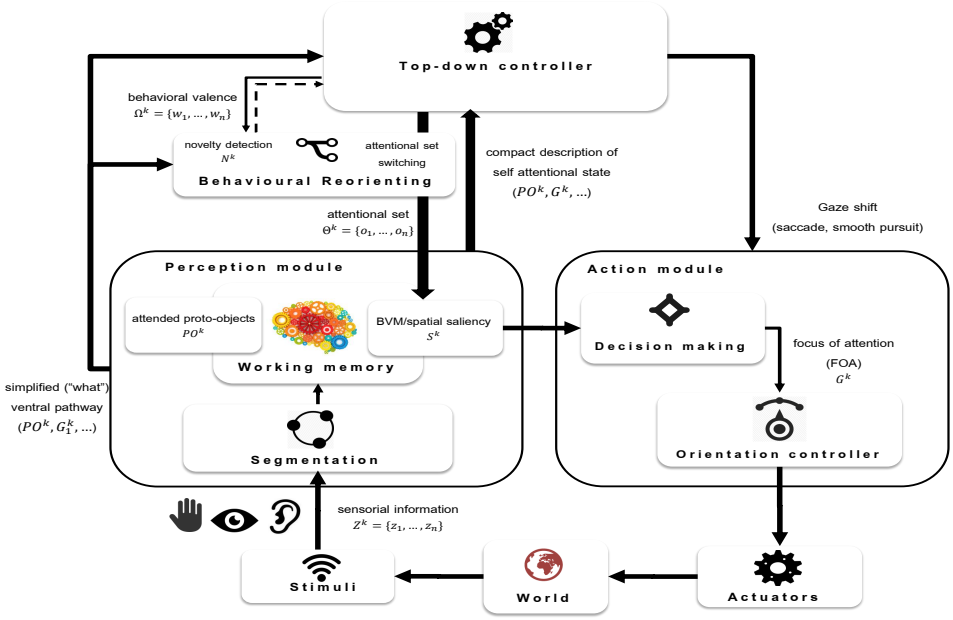
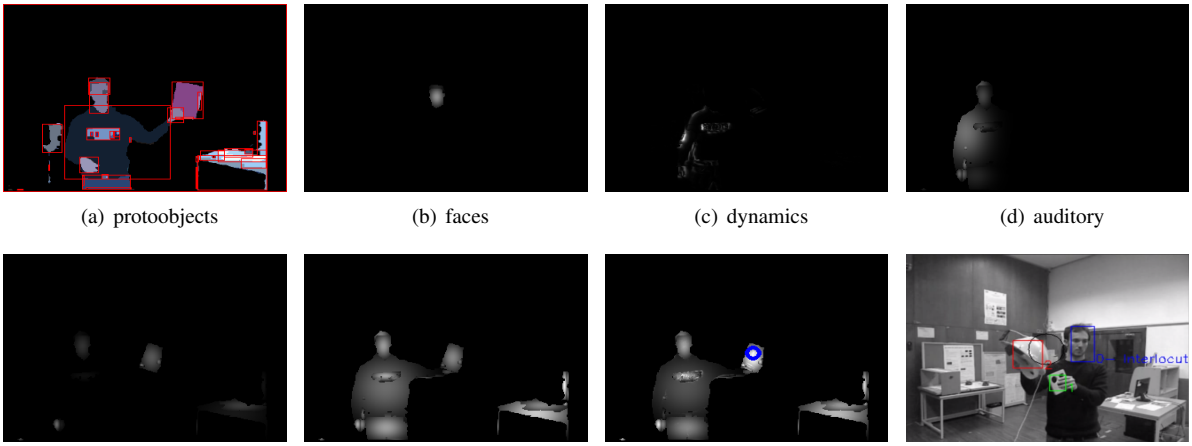


Fig. 1. Attentional system design. The perception module processes sensory signals to build an egocentric representation of the environment (i.e. a spatial saliency map, and a list of spatially-indexed proto-objects) and maintains it in working memory. The top-down controller generates, according to current goals, control signals and sets of relative weights that modulate responses to different features (i.e. the attentional set and behavioural valences). The action module sends commands to actuators according to the attentional map and the gaze shift behaviour informed by the top-down controller. The reorienting module checks for unexpected and behaviourally relevant stimuli, overriding the current attentional set if necessary.



(a) protoobjects (b) faces (c) dynamics (d) auditory

(e) colour bias (f) intensity (g) final saliency (h) memory

Fig. 2. Perception module. (a) shows proto-object (PO) segmentation. Each PO is represented using its average colour (bounding boxes are also plotted for better visualisation); (b, c, d, e, f) show different features associated with the POs (colour contrast, which is not shown, is also used). The top-down modulation will modify the importance of the features as well as the colour bias (i.e., in this case the one used is the pure red); (g) is the final 2D saliency map and the selected PO (blue circle); finally, (h) shows POs (coloured rectangles) stored in working memory by means of deictic pointers.

perception module, which takes input signals provided by sensors and constructs an egocentric representation of the perceived environment that will in turn serve to select the next focus of attention (FOA) according to relevance encoded as saliency; (2) the *top-down controller*, which ensures that the next selected FOA will be influenced by current goals and context; (3) the *action module*, that selects the next fixation location by deciding based on the input from the perception module and provides the control signals to the actuators, according to the current exploration behaviour (i.e. the type of gaze shift strategy, for example, smooth pursuit or saccade generation); (4) the *behavioural reorienting module*, that is in charge of detecting novel and behaviourally-relevant stimuli that should result in interrupting and resetting the attentional process as an action-perception loop.

A. Perception module

This module incorporates working memory that stores two different types of information: a list of attended proto-objects, using a solution similar to [22], and a 3D log-spherical inference map associating saliency to occupied spatial locations developed in previous work [29], [30], [31].

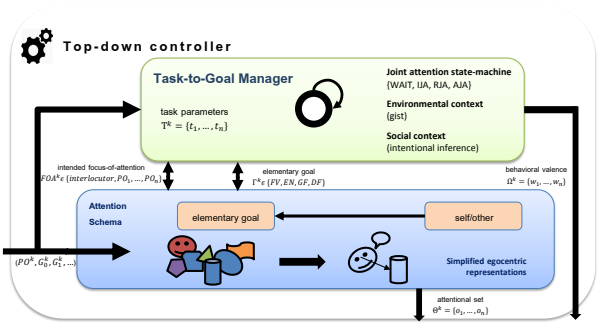


Fig. 3. Top-down controller. The task management module manages high level information about the process, and includes joint attention state machine and a contextual information manager. The attention schema, a simplified model of an underlying attentional process, is in charge of infering/predicting/deciding on the current/next focus of attention and abstract goals of the robot and its interlocutors – refer to main text for more details.

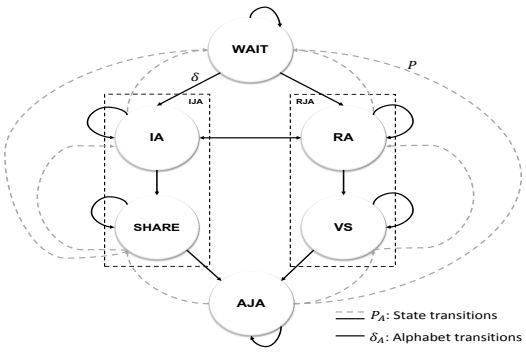


Fig. 4. Joint Attention State Machine. The agent switch between states driven by its own state transition probabilities (P), the possible transitions given the input alphabet (δ) and the current observation of other's state (O).

A preliminary processing stage segments sensor information into pre-attentional volatile perceptual units called *proto-objects* [3]. Feature contrasts are weighted to form the final saliency map S^k at instant k by means of the so-called *attentional set* [11], represented as Θ. This set is provided by the top-down controller, thereby modulating what would be a stimulus-driven process by influence of current goals and context. The sensors observe the world providing the signals Z, which are then transformed into spatial conceptual features F that are filtered by the proto-object set PO and fused into the saliency map S,

$$\mathcal{Z} \rightarrow_\Theta PO \rightarrow_\Theta \mathcal{F} \rightarrow_\Theta \mathcal{S}, \tag{1}$$

representing the relevance of a specific region in space. Each proto-object (PO) is defined as a subset of similar and connected pixels, and its saliency for a specific feature $f \in F$ is defined by a bivariate normal density function $\sim N(\mu = PO_{\text{centre-of-mass}}, \Sigma = diag([PO_{height}, PO_{width}]))$. This ensures that fixations will be drawn to the centre of the PO, as has been proven to generally happen with human attention [32]. Figure 2 shows a few examples of outputs generated by the perception module, from PO segmentation (Fig. 2(a)) to deictic pointers to POs in working memory (Fig. 2(h)).

Proto-objects and their respective deictic pointers, $PO^k = \{PO_1, \cdots, PO_N\}$ are stored in working memory, since they have been (and may be in the near future) FOA. These pointers, besides storing spatial coordinates, associate each PO to its characteristic properties, namely those related to saliency features (e.g. colour). According to neurological studies, humans can keep covert attention (i.e. track without needing to fixate) on 5 objects simultaneously [16].

Stored proto-object information as well as the estimated gaze direction of potential interlocutors $G^k = \{G_1, \cdots, G_M\}$ are provided to the top-down controller with other useful information to allow inference of the own's state and the other's intention.

B. Top-down controller

The top-down controller is depicted in Figure 3 with two representative layers: (1) the task-to-goal manager that is the core of high-level decision making, and (2) the simplified representation of the attentional state in egocentric representation called the "attention schema", as defined in section II. We consider each interlocutor agent (human/robot) as entities capable of knowing and predicting their own internal state and estimating the other's state. Therefore the task-to-goal manager is basically in charge of: generating own state according to current goals and an estimate of the other's state by means of a probabilistic state machine, and estimating the other's state by using the set of high-level signals that describe other's hidden process. The module outputs the set of parameters that modulates the attentional process and the behavioural valences that define the importance of unexpected stimuli, according to current goals and task.

For generating the robot's own state, we define the Joint Attention State Machine (JASM) as an extension of the probabilistic finite state automata[1][8], for which the input sequence of the alphabet is the predicted other's state. The JASM is described by (see Fig. 4)

$$\mathcal{A} =< Q, \Sigma_A, \Sigma_\Gamma, \delta, I, P, O, \Omega, \Gamma >, \tag{2}$$

Notation is as follows:

- Q = {WAIT, IA, SHARE, RA, VS, AJA} - set of states in the joint attention process. WAIT represents that the agent is not interacting but is waiting for a signal input (e.g., a human not engaging but passing through the social space). IA and SHARE are two states derived from the IJA process due to differences on their attentional parameters values. The former represents any type of engaging or initiating the interaction while the latter corresponds to the action of sharing an object with the other. RA and VS (Visual Search) are two substages of the RJA process, the former describe the

[1]It differs from the standard probabilistic automata because the appearance of the alphabet symbols is an observation process subject to uncertainty and there are not final probabilities. Besides, as the signals are outputted when the automata is in a specific state, it can also be considered an extension of a hidden Markov model [8].

initial response to other's IJA and the latter is the action of searching the object that the other want to share. Finally, AJA is the last stage of the triadic relation where the agent communicates that understands that the sharing is complete (e.g., engaging the other after the VS or SHARE state). It is important to highlight that while WAIT and AJA can be overlapped in both agents, when one of them is in the set (IA, SHARE) the other should correspond with the set (RA, VS).

- Σ_A = {WAIT, IA, SHARE, RA, VS, AJA} - alphabet accepted by the automata that matches the other's state.
- Σ_Γ = {FV,EN,GF,DF} - alphabet of the emissions produced in each state defined as the high level action to be performed. FV means free view of the agent; EN represents engaging the other; GF describes gaze following; and DF means deictic fixation of an object.
- $O \in \Sigma_A$ - observations by estimating other's state.
- $\delta \subseteq Q \times \Sigma_A \times Q$ - transitions depending on the alphabet.
- $I : Q \to \mathbb{R}^+$ - initial state probabilities.
- $P : \delta \to \mathbb{R}^+$ - transition probabilities given the input alphabet.
- $\Omega : O \times Q \to \mathbb{R}^+$ - set of conditional observation probabilities.
- $\Gamma : \Sigma_\Gamma \times Q \to \mathbb{R}^+$ - state emission probability function.

The generative nature of the probabilistic automata is used as the central core of the top-down controller that will switch states for itself and depending on the other's state. Given the variable ${}^AQ_i^k$, the pre-superscript A represents the agent and k denotes the instant, and the subscript i indexes the state in the set. In order to model how much time an agent can remain in one state we include a random variable T that has an exponential density function $P(T_i^k|Q_j^k) = exp(-\lambda^k)$, where $\lambda \in [0, 1]$. This makes that the probability of being in a particular state decreases with the time, forcing the agent to switch between states.

We solve the automata state selection by Maximum A Posteriori (MAP) estimation over a Bayesian filter. Thus, in order to select the current own state (${}^AQ^k$) given the observation of the other state (${}^BQ^k$) we use the following question:

$$P({}^AQ^k|O^k, {}^BQ^k, {}^BT^k) \simeq$$
$$\simeq P(O^k|{}^BQ^k)P({}^AT^k|{}^AQ^k)\sum_{i\in Q} P({}^AQ_i^k|{}^AQ_i^{k-1}) \quad (3)$$

We model this hidden process given the attentional cues and the observed other's goals by a dynamic Bayesian network (i.e. an adapted hidden Markov model) where the observations are given by soft evidence. These observations describe the nature of the other's attention and taking into account that transiting from one state to another is affected by own's actions, we can estimate the current other's state by means of its observed emissions ${}^B\Gamma \in \{FV, EN, GF, DF\}$, and the current own's goals emissions ${}^A\Gamma$. Thus, other's state estimation is updated by O^k:

$$P({}^BQ^k|{}^B\Gamma^k) \simeq P(O^k|{}^BQ^k)P({}^BQ^k) \quad (4)$$

The dynamics of the process is captured by means of the probability of the other's transiting from one state to another given the probability of remaining in that state when our own state is emitting a particular signal Γ:

$$P({}^BQ^{k+1}|{}^BQ^k, {}^A\Gamma^k) =$$
$$= \sum_{j\in Q} P({}^BQ^{k+1}|{}^BQ^k, {}^A\Gamma^k)P({}^BT^k|{}^BQ^k)P({}^BQ^k) \quad (5)$$

Note that we need to estimate or learn the forward model defined by $P({}^BQ^{k+1}|{}^BQ^k, {}^A\Gamma^k)$ experimentally [33].

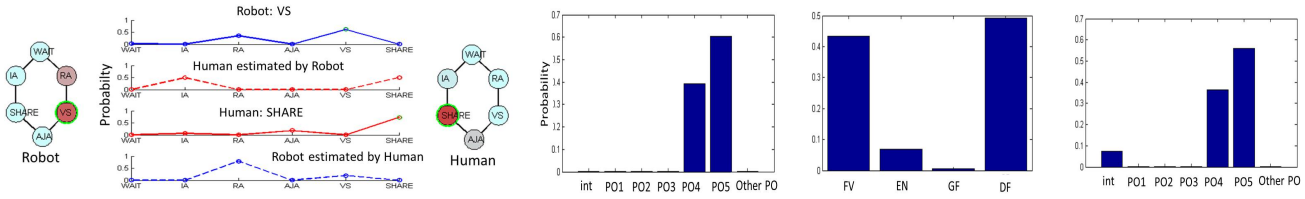
Following the awareness model of the human brain (see section II) we designed a simplified representation attention, similar to a framework proposed by Gilet et al. for handwriting analysis and reproduction [34]. The attentional cues that arrive from the perception module are used to infer the intended focus of attention and goals of the other by means of the self goals provided by the task-to-goal manager in the form of JASM emissions Γ^k, and also to predict the consequences of the robot's next FOA according to current goals and as such select the next parameter set that will modulate attention. This simplified model provides the needed abstraction from the complications of the underlying attentional processes, in a very tractable yet effective fashion. The FOA of the self or the other are referred to in this model in the robot's egocentric point-of-view, thereby integrating spatial cues into a common reference. The predictive trait of this model resembles the *efferent copy* mechanism of the human brain enacted by mirror neurons [34], [3].

C. Behavioural Reorienting

This module is in charge of overriding the attentional set when an unexpected stimulus with behavioural relevance is sensed, therefore resetting the attentional process. The behavioural valence modulates the importance of the different stimuli in face of current context and goals, as imposed by the top-down controller. For instance, most auditory onsets should not distract the robot from attention-demanding tasks such as engaging with the current interlocutor; however, a sudden/loud/unexpected noise, especially if coming from outside of the field of view, should promote breaking the robot's concentration so as to enable it to attend to a potential danger. Novelty can be computed using Bayesian surprise theory [20] to analyse the importance of changes in the distributions of the 3D inference map in two consecutive instants.

D. Action module

This module is in charge of deciding the final FOA and the best control actions to attend the specified location taking into account the current agent state. For the orientation controller we distinguish two different modes of operation: saccadic behaviour, in which the robot performs a quick gaze shift to the desired FOA; and smooth pursuit, in which the robot smoothly tracks the current object of interest, making it a persistent FOA. As the perceptual representations of the system are in egocentric coordinates, the orientation module includes a feedback controller that uses as input the current



(a) JASM (b) self FOA given goal=GF (c) other's goal (d) other's FOA

Fig. 5. Top-down controller simulation. (a) The robot (at the JASM) is switching towards visual search (VS) because the human has transited from IA to SHARE phase. The estimated distribution of other's state still reflects the transition. The continuous / discontinuous lines show own state and other's estimation respectively. (b,c,d), show the corresponding intentions inferred by the attentional schema.

FOA and as output the control signals for the actuators. During saccadic behaviour, next FOA selection is performed through selecting the location with maximum saliency in a process similar to what is described in [31], while for smooth pursuit the fixation location for the current FOA is computed by adding to this process a "sharp" probability distribution centred on the object of interest, therefore promoting fixations on regions of high saliency in close proximity to the tracked object.

V. Results

A robotic attentional system must deal with real-time processing of sensory signals, as well as the integration and synchronisation of several state-of-the-art components. In our case, this means correctly dealing with signal segmentation, 3D egocentric saliency representation [29], gaze inference (e.g., head pose inference and pupil detection), working memory management, FOA selection, saccade control, top-down modulation (own and other's state estimation), etc. In this paper, we report on the current implementation of the perception and action modules through experimental online validation, and on the top-down controller in simulation.

We are currently working on the definitive and complete version of the proposed attentional system implementation, for which the main missing link is currently the gaze inference module. Preliminary work on robust Bayesian gaze estimation has just been finished, exhibiting satisfactory and robust performance, and we are currently concluding a real-time implementation.

The current implementation, developed using the Robotic Operating System (ROS), operates at 12 fps for PO segmentation, 8 fps for saliency computation, and from 8 to 20 fps (when tracking a PO) for working memory management. This means that we can achieve the same performance as the human saccade-generation system – just under 500 ms on average between fixations [3]. The top-down controller timing is non-critical and therefore computational time analysis is not needed. Additionally, the auditory saliency is computed by using open source robot audition system HARK [35], the reorientation module currently only takes into account auditory signals, and the PO tracker is an adapted version of [36] for multiple objects. A detailed specification of the robot head and its sensors can be found in [37].

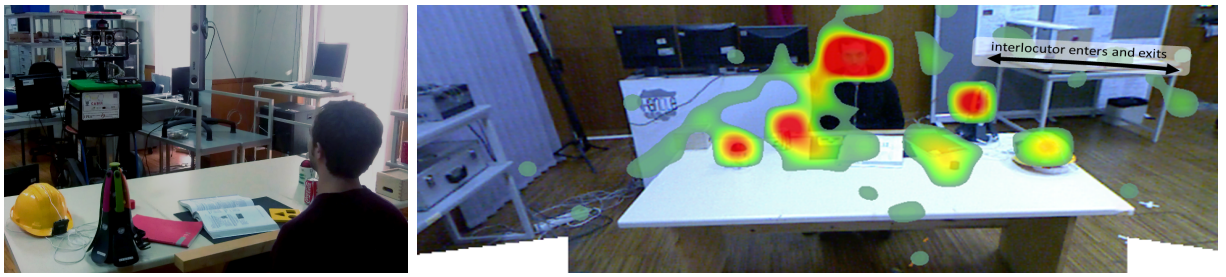
A. Top-down module

We have tested the JASM by simulating the interlocutor intentional state, and analysing its outputs. Results show that the robot is able to behave coherently with the other's intentional state (the mirroring response is 67% and the number of completed joint attention tasks when the human initiates and completes the behaviour is 78.40%) and even to spontaneously initiate joint attention (the 47.21% of the total IAs is performed by the robot for a 10000 transitions simulation) – see Figure 5(a). On the other hand, by simulating the signals provided by the perception module, we show how the attentional schema computes the intention probabilities of own and other's state (those output distributions feed the JASM – see Figure 5(b) and 5(c)). In the example depicted in the figure, the robot is following the interlocutor's gaze, and he/she is inferred to most probably be intentionally looking at an object (i.e. DF state). Therefore, the most probable PO to fixate is the most salient PO within the interlocutor's line-of-sight, in this case PO_5, in that moment already stored in working memory.

B. Perception and action modules

The experimental set-up used to test these modules in realistic conditions is depicted in Fig. 6(a), where an interlocutor is in front of a set of distinct objects over a table. First we evaluated the overt attention system response in free view, and then we emulated a simple behaviour of the top-down controller that would promote the following sequence of events: (1) the system is in free view until it discovers the interlocutor; (2) the interlocutor shows an object to the robot; (3) the robot acknowledges the object and set its colour as a bias for perception; (4) the robot performs a visual search until it finds an object with similar characteristics. It is important to highlight that the system will only use indirect colour bias modulation and tracking to onset these events. To perform the statistical evaluation of the interaction behaviour, we record several individuals interacting with the robot and then classify the behaviours of both according to their respective reactions.

Figure 6(b) shows a stitched image of the scene with a visual attention heatmap of the free view experiment superimposed. The most attended locations correspond to the interlocutor's face and objects with a high red colour component (i.e. in this way, we model human phylogenetic



(a) set-up (b) attention heatmap

(c) state 0 (d) state 1 (e) state 2

Fig. 6. Perception experiments. (a) the set-up is compound by a robot provided with an active head in front of a table with a set of objects. (b) attentional map in free view. Study case experiment sequence: (b) the robot is in free view where it has added an object and later detected the interlocutor; (c) afterwards, the human shows an interesting object to the robot, that it is acknowledged by the robot. (d) the object colour is used as a bias to initiate the visual search until it finds one similar object. Correct acknowledgement of the task is demonstrated if the fixated object matches expectations.

bias towards red). Yellow objects also attract robot's attention due to proximity to red in colour space and their relatively high intensity contrast. Figures 6(c), 6(d) and 6(e) show an experiment using our simple model of "intelligence" to test the attentional system. The robot performs an interesting and coherent behaviour despite of the reactive underpinning of the perception and action modules.

Finally, we evaluated reciprocal human-robot interaction [38] by analysing the robot's expected behaviour when faced with different individuals. For each trial the interlocutor is asked to pick up a red, yellow or blue object. The evaluation scenario, which although relatively controlled is already open-ended and challenging, can be characterised as follows: (1) the system had no internal prior expectations, neither over the objects nor the interlocutors; (2) interaction could occur anywhere within 1 to 4 metres from the robot; (3) apart from the general task, no other indication or scripting, spatial or temporal, was given to the interlocutor.

Table I shows the number of times fixation behaviour occurred as expected given the total of key fixation instants. Expectations were considered to be met whenever the system was deemed by visual assessment to be enacting the behaviours labelled in the top row of the table, in correct order. A high percentage of success was found in engagement, visual search and acknowledgement. Conversely, a low success rate was found in shifting gaze towards the interlocutor's FOA result mainly from the lack of gaze inference and gist modulation. The low realization in VS at the "red" colour is due to the similar response of the saliency to yellow (i.e., after biasing to "red", sometimes, the next selected object was yellow).

TABLE I

ANALYSIS OF EXPECTED ROBOT BEHAVIOUR

Trial conditions	Engage %	Fixate Object %	Visual Search %	Acknowledge %
Red	76.81	34.30	42.15	65.10
Yellow	79.00	50.15	60.14	76.66
Blue	73.50	22.54	55.88	66.13
Total	76.44	35.66	52.72	69.30

VI. DISCUSSION

We have presented an overarching framework implementing artificial attention, designed to fulfil the requirements of social interaction (i.e. reciprocity and awareness), with strong inspiration on current theories in functional neuroscience, described in section II.

The emergence of an inkling of intelligent behaviour due to the interconnection of multiple independent elements, even in its current open-loop operation mode, has shown the potential of the perception and action modules. The top-down controller has been shown to operate as expected under simulation, suggesting that, indeed, system behaviour will be significantly improved when the perception and action modules become ready to be modulated by top-down influences. This will introduce meaningful repeatability, and consequently the expectation of the interlocutor can be effectively fulfilled. Moreover, we believe that the JASM and the attentional schema offer exciting new insights on how non-deterministic probability states machines that could give the robot a more conceptual sense of adaptive behaviour and even free will. Additionally, a great challenge is involved in correctly learning the actual transition probabilities using human interaction data. In terms of the action module, although

the FOA selection using a heuristic function seems to work as expected, approaches such as decision-making methods for autonomous agents perception and control [39], could be interesting in order to maximize the obtained information during visual search. Finally, we are currently designing an experimental paradigm to use this system to evaluate the influence of attention on HRI, the foundation of which is based on the already published methodologies of [6], [40].

REFERENCES

[1] M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino, and C. Yoshida, "Cognitive developmental robotics: a survey," *Autonomous Mental Development, IEEE Transactions on*, vol. 1, no. 1, pp. 12–34, 2009.

[2] B. Scassellati, "Theory of mind for a humanoid robot," *Autonomous Robots*, vol. 12, no. 1, pp. 13–24, 2002.

[3] J. F. Ferreira and J. Dias, "Attentional Mechanisms for Socially Interactive Robots – A Survey," *IEEE Transactions on Autonomous Mental Development*, vol. 6, no. 2, pp. 110–125, 2014.

[4] C. Breazeal, "Toward sociable robots," *Robotics and Autonomous Systems*, vol. 42, no. 3–4, pp. 167–175, 2003.

[5] B. Scassellati, "Investigating models of social development using a humanoid robot," in *Neural Networks, 2003. Proceedings of the International Joint Conference on*, vol. 4, 2003, pp. 2704–2709 vol.4.

[6] P. Lanillos, J. F. Ferreira, and J. Dias, "Evaluating the influence of automatic attentional mechanisms in human-robot interaction," in *Workshop: a bridge between Robotics and Neuroscience Workshop in Human-Robot Interaction, 9th ACM/IEEE International Conference on*, Bielefeld, Germany, March 2014.

[7] J. F. Ferreira and J. Dias, *Probabilistic Approaches to Robotic Perception*. Springer, 2014.

[8] E. Vidal, F. Thollard, C. De La Higuera, F. Casacuberta, and R. C. Carrasco, "Probabilistic finite-state machines-part i," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 7, pp. 1013–1025, 2005.

[9] J. Hohwy, "Attention and conscious perception in the hypothesis testing brain," *Frontiers in Psychology*, vol. 3, no. 96, 2012.

[10] C. Lord, S. Risi, L. Lambrecht, E. H. Cook Jr, B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter, "The autism diagnostic observation schedule-generic: A standard measure of social and communication deficits associated with the spectrum of autism," *Journal of autism and developmental disorders*, vol. 30, no. 3, pp. 205–223, 2000.

[11] M. Corbetta and G. L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nature reviews neuroscience*, vol. 3, no. 3, pp. 201–215, 2002.

[12] M. Corbetta, G. Patel, and G. L. Shulman, "The reorienting system of the human brain: from environment to theory of mind," *Neuron*, vol. 58, no. 3, pp. 306–324, 2008.

[13] S. Vossel, J. J. Geng, and G. R. Fink, "Dorsal and ventral attention systems distinct neural circuits but collaborative roles," *The Neuroscientist*, vol. 20, no. 2, pp. 150–159, 2014.

[14] M. S. Graziano, *Consciousness and the social brain*. Oxford University Press, 2013.

[15] M. S. Graziano and S. Kastner, "Human consciousness and its relationship to social neuroscience: a novel hypothesis," *Cognitive neuroscience*, vol. 2, no. 2, pp. 98–113, 2011.

[16] M. J. Spivey, D. C. Richardson, and S. A. Fitneva, "Thinking outside the brain: Spatial indices to visual and linguistic information," *The interface of language, vision, and action: Eye movements and the visual world*, pp. 161–189, 2004.

[17] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[18] A. Garcia-Diaz, V. Leborán, X. R. Fdez-Vidal, and X. M. Pardo, "On the relationship between optical variability, visual saliency, and eye fixations: A computational approach," *Journal of vision*, vol. 12, no. 6, p. 17, 2012.

[19] N. D. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *Journal of vision*, vol. 9, no. 3, p. 5, 2009.

[20] P. Baldi and L. Itti, "Of bits and wows: a bayesian theory of surprise with applications to attention," *Neural Networks*, vol. 23, no. 5, pp. 649–666, 2010.

[21] J. Tünnermann, C. Born, and B. Mertsching, "Top-down visual attention with complex templates." in *VISAPP (1)*, 2013, pp. 370–377.

[22] A. J. Palomino, R. Marfil, J. P. Bandera, and A. Bandera, "Multi-feature bottom-up processing and top-down selection for an object-based visual attention model," in *2nd Workshop on Recognition and Action for Scene Understanding (REACTS)*, 2013.

[23] M. Aziz and B. Mertsching, "Visual search in static and dynamic scenes using fine-grain top-down visual attention," in *Computer Vision Systems*, ser. Lecture Notes in Computer Science, A. Gasteratos, M. Vincze, and J. Tsotsos, Eds. Springer Berlin Heidelberg, 2008, vol. 5008, pp. 3–12.

[24] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Progress in brain research*, vol. 155, pp. 23–36, 2006.

[25] F. Shic and B. Scassellati, "A behavioral analysis of computational models of visual attention," *International Journal of Computer Vision*, vol. 73, no. 2, pp. 159–177, 2007.

[26] B. Kuhn, B. Schauerte, K. Kroschel, and R. Stiefelhagen, "Multimodal saliency-based attention: A lazy robot's approach," in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*. IEEE, 2012, pp. 807–814.

[27] G. Metta, G. Sandini, and J. Konczak, "A developmental approach to visually-guided reaching in artificial systems," *Neural Networks*, vol. 12, no. 10, pp. 1413 – 1427, 1999.

[28] D. Vernon, G. Metta, and G. Sandini, "A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents," *Evolutionary Computation, IEEE Transactions on*, vol. 11, no. 2, pp. 151–180, 2007.

[29] P. Lanillos, J. F. Ferreira, and J. Dias, "Multisensory 3D Saliency for Artificial Attention Systems," in *3rd Workshop on Recognition and Action for Scene Understanding (REACTS)*, 2015.

[30] J. F. Ferreira, J. Lobo, P. Bessière, M. Castelo-Branco, and J. Dias, "A Bayesian Framework for Active Artificial Perception," *IEEE Transactions on Cybernetics (Systems Man and Cybernetics, part B)*, vol. 43, no. 2, pp. 699–711, April 2013.

[31] J. F. Ferreira, M. Castelo-Branco, and J. Dias, "A hierarchical Bayesian framework for multimodal active perception," *Adaptive Behavior*, vol. 20, no. 3, pp. 172–190, June 2012.

[32] V. Yanulevskaya, J. Uijlings, J.-M. Geusebroek, N. Sebe, and A. Smeulders, "A proto-object-based computational model for visual saliency," *Journal of Vision*, vol. 13, no. 13, p. 27, Nov. 2013.

[33] A. P. Shon, J. J. Storz, and R. P. Rao, "Towards a real-time bayesian imitation system for a humanoid robot," in *Robotics and Automation, 2007 IEEE International Conference on*. IEEE, 2007, pp. 2847–2852.

[34] E. Gilet, J. Diard, and P. Bessière, "Bayesian action–perception computational model: interaction of production and recognition of cursive letters," *PloS one*, vol. 6, no. 6, p. e20387, 2011.

[35] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Design and implementation of robot audition system "hark" – open source software for listening to three simultaneous speakers," *Advanced Robotics*, vol. 24, no. 5-6, pp. 739–761, 2010.

[36] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, 2014.

[37] P. Lanillos, J. Oliveira, and J. F. Ferreira, "Experimental Setup and Configuration for Joint Attention in CASIR," ISR, Coimbra, Tech. Rep., 2013. [Online]. Available: http://mrl.isr.uc.pt/projects/casir/files/MRL-CASIR-TR-2013-11-TR001.pdf

[38] A. L. Thomaz and C. Breazeal, "Experiments in socially guided exploration: Lessons learned in building robots that learn with and without human teachers," *Connection Science*, vol. 20, no. 2-3, pp. 91–110, 2008.

[39] P. Lanillos, S. K. Gan, E. Besada-Portas, G. Pajares, and S. Sukkarieh, "Multi-UAV target search using decentralized gradient-based negotiation with expected observation," *Information Sciences*, vol. 282, no. 0, pp. 92 – 110, 2014.

[40] G. Avraham, I. Nisky, H. L. Fernandes, D. E. Acuna, K. P. Kording, G. E. Loeb, and A. Karniel, "Toward perceiving robots as humans: Three handshake models face the turing-like handshake test," *Haptics, IEEE Transactions on*, vol. 5, no. 3, pp. 196–207, 2012.