Multisensory 3D Saliency for Artificial Attention Systems

Pablo Lanillos[1] and João Filipe Ferreira[1] and Jorge Dias[12]

[1] AP4ISR team, Institute of Systems and Robotics (ISR) Dept. of Electrical & Computer Eng., University of Coimbra. Pinhal de Marrocos, Polo II, 3030-290 COIMBRA, Portugal

[2] Khalifa University of Science, Technology, and Research Abu Dhabi 127788, UAE

Abstract. In this paper we present proof-of-concept for a novel solution consisting of a short-term 3D memory for artificial attention systems, loosely inspired in perceptual processes believed to be implemented in the human brain. Our solution supports the implementation of multisensory perception and stimulus-driven processes of attention. For this purpose, it provides (1) knowledge persistence with temporal coherence tackling potential salient regions outside the field of view, via a panoramic, log-spherical inference grid; (2) prediction, by using estimates of local 3D velocity to anticipate the effect of scene dynamics; (3) spatial correspondence between volumetric cells potentially occupied by proto-objects and their corresponding multisensory saliency scores. Visual and auditory signals are processed to extract features that are then filtered by a proto-object segmentation module that employs colour and depth as discriminatory traits. We consider as features, apart from the commonly used colour and intensity contrast, colour bias, the presence of faces, scene dynamics and also loud auditory sources. Combining conspicuity maps derived from these features we obtain a 2D saliency map, which is then processed using the probability of occupancy in the scene to construct the final 3D saliency map as an additional layer of the Bayesian Volumetric Map (BVM) inference grid.

1 Introduction

Many of the cognitive skills that animals display are strongly dependent on attention. For example, toddlers, at 18 months of age, reach the representational stage [4, 20], at which they are able to interpret the caregiver's gaze and reorient their head towards a potentially shared location, thus enacting the basic joint action called *joint attention* [4]. In fact, one of the key skills in humans is the ability to discern what are the important locations to pay attention according to the environmental and social cues. Recent neurophysiology studies [2, 24] reinforce the idea that two intertwined streams of attentional processing exist in the human brain. One of them, commonly referred to as bottom-up (stimulus-driven) attention, has been computationally modelled through saliency approaches [4]. In this respect, the seminal work of Itti et Al. [9] presents a method to compute

the saliency map using only images (i.e., visual stimuli). Several features of the image (colour, orientation) are processed to form conspicuity maps, which are subsequently fused into a final image denominated *saliency map*, attributing priority values to each pixel. In other words, each pixel has an associated saliency value that determines the *behavioural relevance* of that location, in order to be able to decide the next focus of attention. However, enabling a robot to engage in social interaction by means of tasks such as joint attention, which is not merely a process of coordinating gaze, is particularly challenging. In fact, behavioural relevance scores are needed to interpret the intention, which implies more than just a simple, stimulus-driven process. This means that a way of modulating bottom-up attention by goal-oriented signals is necessary, and hence implementing a complementary top-down modulated attentional stream, as identified in [2, 24], is as relevant for artificial perception systems as it is for the human brain.

Attention is in essence a dynamic and multisensory process [4]. Be it in social interaction, where both verbal and non-verbal signals must be singled out from all the additional irrelevant stimuli, or even in other common tasks such as driving, visual cues alone are not enough for correct and adaptive behaviour. Moreover, attention is known to involve objects outside of the field of view, a fact that has been at most dealt in the literature by means of heuristic approaches [4].

In the brain, it is mostly consensual that, although within the preattentive stages objects are not fully recognised, some features are grouped into larger attention units called proto-objects. However, how proto-objects are handled within attention and their spatial correspondence in the working memory is still a matter of debate. In fact, many of the current approaches first compute the saliency and afterwards try to segment the different proto-objects [8], losing primordial information provided by the sensors. Conversely, we believe there are definite benefits in segmenting proto-objects first in an emulation of the preattentive stage of the human brain and only consequently attributing saliency values to each proto-object. As in [17] and [25], we propose a bottom-up approach framed into a full artificial attention system that allows to evaluate the saliency map and the potential attended object at the same time from visual and auditory cues. Moreover, besides solving the difficulties of going from pixel-wise saliency to proto-objects, we also incorporate a way, inspired in the work presented in [21], to ensure that fixations will be drawn to the centre of the proto-object as recent studies in humans point out, therefore eliminating the bias towards object boundaries introduced by most solutions in the literature [25].

In this paper we present proof-of-concept for a novel solution consisting of a short-term 3D memory for artificial attention systems, loosely inspired in perceptual processes believed to be implemented in the human brain. Our approach is an extension and improvement of previous work including an inference grid associated to an amodal (i.e. non-sensor specific), egocentric, 3D log-spherical spatial configuration first presented in [3, 5]. Our solution supports the implementation of multisensory perception and stimulus-driven processes of attention, according to a bioinspired view stemming from current research from psychology

and neuroscience (e.g. [18]). It states that stochastic models of the environment, which are continuously updated on the basis of processed sensory information, might exist in the brain. Our solution provides (1) knowledge persistence with temporal coherence tackling potential salient regions outside the field of view, via the panoramic, log-spherical inference grid; (2) prediction, by using estimates of local 3D velocity to anticipate the effect of scene dynamics; (3) spatial correspondence between volumetric cells potentially occupied by proto-objects and their corresponding multisensory saliency scores.

The remainder of this paper is structured as follows. Section 2 discusses different computational approaches for bottom-up attention and the main motivation of this work. Section 3 explains the different stages to compute the proto-object based 3D saliency. Section 4 analyse experimentally the proposed approach and describes in detail its implementation. Finally, section 5 summarizes the proof-of-concept results, comments the 3D saliency advantages observed and sketches the future work.

2 Computational Bottom-up Attention Approaches

Several approaches in modelling bottom-up attention have been proposed in the literature. The majority are based on [9], and propose the computation of pixelwise 2D image saliency based on feature contrast. A strong competitor is the work by Tsotsos et al. (e.g. [1]), that uses the information gain that the system obtains when attending a specific location, being mostly applied to visual search [7]. Entropy and information maximization approaches have been successfully applied in many robotics solutions like attention or decentralized search [12], but they have been found not to be able to solve the whole bottom-up process. In fact, recent research in neuroscience supporting that the attention process is affected by novelty [2] suggests that information gain should be used inside the artificial attention system complementing other processes. As opposed to pixelwise approaches, a minimalistic segmentation of the image into proto-objects can also be performed to then extract the full details of the attended object when fixation occurs [19]. We argue that, although attention serves for allocating perceptual resources to analyse a subset of incoming stimuli in detriment of the rest of the perceptual scene [4], some computational resources should nevertheless be devoted in implementing a preattentive stage using fast segmentation algorithms. On one hand, perceptual grouping of colour, depth and other basic characteristics into proto-objects can be achieved by pyramidal segmentation [15] and its computation burden can be tackled using parallelisation. In fact, recent approaches [21, 17, 14] include proto-object segmentation as the core of the saliency system. On the other hand, face detection has been demonstrated to be a fast process (38 ms) in the human brain [13], and has therefore been suggested as a possible basic feature due to phylogenetic reasons [4]. Using it as an important feature to draw attention can make the difference when developing non-verbal skills in HRI. Finally, we can find in the literature a few multisensory 2D bottom-up attention approaches like [20], in which saliency is computed also

using auditory cues. Establishing attention as a multisensory process implies that spatial correspondence is needed [19]. In other words, saliency maps resulting from visual and auditory perception must be both related and integrated into a single egocentric reference [4]. Additionally, saliency maps must be temporally registered in order to deal with egomotion (i.e., head and body motion). Recent research efforts have introduced spherical representations of egocentric space to deal with these issues [22, 5].

Our proposal extends bottom-up 2D saliency approaches to an amodal 3D representation by combining proto-object driven segmentation with multisensory cues in a 3D log-spherical inference grid [5]. Therefore, it provides a saliency representation of the surrounding environment from which the potential objects can be retrieved for further analysis.

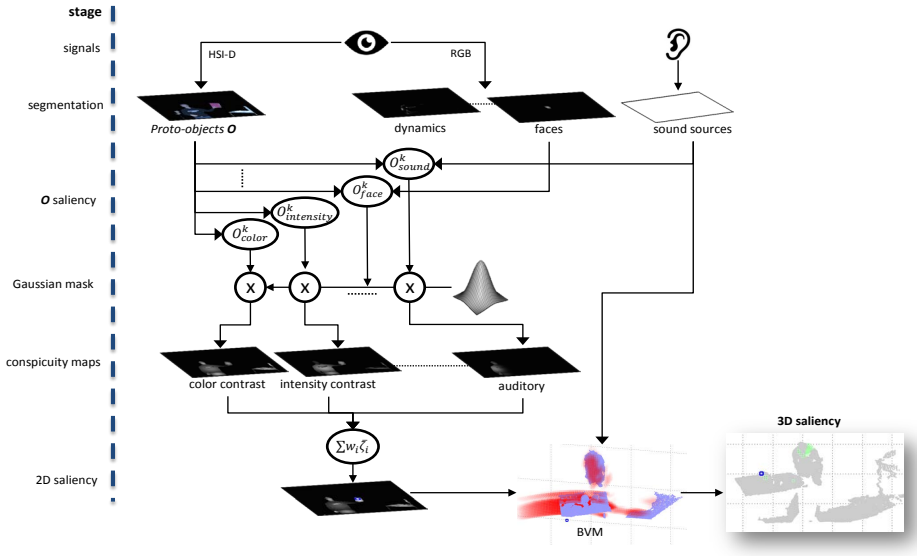


Fig. 1. Process flow: from signals to 3D saliency map.

3 Protoobject-based 3D Saliency

For constructing the attention short-memory we follow the process depicted in Fig. 1. Our system segments visual and auditory cues into different features, and segments proto-objects on the image. Then, each feature is computed for each proto-object and afterwards a set of Gaussian masks produces the conspicuity maps that define the value of the feature at a pixelwise level. The 2D saliency is computed mixing all maps according to their importance. In parallel, a module

implementing the BVM inference grid [5] uses depth data in order to infer the probability of each grid cell being occupied. Finally, the information provided by the 2D saliency map is fused with the BVM by means of a hierarchical Bayesian framework in order to obtain the 3D log-spherical map as if it were a panoramic view of the saliency, taking into account rotational egomotion.

Table 1. Notation

$\mathcal{I}$	image in the HSI space $I_{h\|s\|i}$
$\mathcal{I}(j)$	linear indexed pixel j of the image
$\mathcal{I}(i,j)$	pixel (i,j) of the image
$s(j)$	similarity function abstraction from the BIP segmentation
$\mathcal{O}^k$	proto-object k of the image
$\mathcal{O}^k_{property}$	proto-object k property value
$\bar{\zeta}_{name}$	saliency conspicuity map for feature <name>
ζ^k_{name}	saliency of proto-object k for feature <name>
θ	azimuth angle
θ_{fov}	azimuth field of view of the sensor
$\mathbb{1}_R$	indicator function for subset R
$A \oplus B$	dilation operation to image A using mask B
$dist_{HSI}(\mathcal{O}^k, \mathcal{O}^j)$	distance between proto-object i and j in the HSI space
$adj(\mathcal{O}^k)$	set of proto-objects adjacent to k
$g(\mathcal{O}^k)$	Gaussian function centred in the proto-object k
$\#\text{adjpx}(\mathcal{O}^k, \mathcal{O}^j)$	number of adjacent pixels between proto-object k and j
$\tilde{C}_{ref}$	colour reference in the HSI space
$P(.)$	probability distribution
O_c	occupancy for cell c
$\mathcal{S}$	2D saliency
$\mathcal{S}^t_c$	3D saliency for cell c at instant t
V_c	velocity at cell c
A_c	antecedent of cell c (i.e., previous cell)

3.1 2D saliency

Following the notation defined in Table 1, saliency is computed by combining different features segmentation with an adapted version of the Bounded Irregular Pyramid (BIP) [15] to extract the proto-objects. We construct each $\mathcal{O}$ as the union of the pixels that satisfies the BIP grouping similarity process:

$$\mathcal{O}^k = \bigcup_{j|s(\mathcal{I}(j))\leq\epsilon} j \tag{1}$$

Every $\mathcal{O}^k$ is defined as set of properties: hue (H), saturation (S), intensity (I), perimeter, area, width and height. We distinguish the 2D saliency for each proto-object k and for each feature (ζ^k_{name}) from the conspicuity map that represents

the saliency at a pixel level ($\overline{\zeta}_i$). First, all saliency features (ζ^k_{name}) are computed as follows:

Colour contrast (CC). It defines the colour difference between the proto-object and its neighbours (i.e., proto-objects that are in contact at pixel level in the image) [14]. This feature helps to pop-up objects that are attractive according to its local surrounding.

$$\zeta^k_{CC} = \frac{\mathcal{O}^k_S}{\mathcal{O}^k_{perimeter}} \sum_{j \in \mathcal{O}} dist_{HSI}(\mathcal{O}^k, \mathcal{O}^j) \tag{2}$$

then it is normalized dividing by $\max_k \zeta^k_{CC}$.

Intensity contrast (IC). It expresses, analogously to colour contrast, the intensity component difference of the proto-object according to its neighbours [14]. Yellow objects over black background or lights will stand above with this feature.

$$\zeta^k_{IC} = \frac{1}{\mathcal{O}^k_{perimeter}} \sum_{j \in \mathcal{O}} \Big(|\mathcal{O}^k_I - \mathcal{O}^j_I| \cdot \#\text{adjpx}(\mathcal{O}^k, \mathcal{O}^j) \Big) \tag{3}$$

then it is normalized dividing by $\max_k \zeta^k_{IC}$.

Face detection (face). There is quite a bit of controversy regarding the features that should be computed during the preattentive stage [4]; however, as explained in section 2, there is some support that a dedicated system exists in the brain for detecting faces at this stage. This justifies the idea of modelling a fast detection system that helps the attention process to evaluate the scene. Given a region R defined by the bounding box of a detected face we can compute the face saliency of a proto-object as,

$$\zeta^k_{face} = \frac{1}{\mathcal{O}^k_{area}} \sum_{j \in \mathcal{O}^k} \mathbb{1}_R \tag{4}$$

Auditory source (sound). In our proposed solution, knowing the position of an auditory source emitting sound louder than a predefined threshold currently contributes to 2D saliency by augmenting the importance of co-localised proto-objects. For instance, a "face that is talking" should be more relevant than other objects. We model the proto-object saliency as follows:

$$\begin{aligned} \zeta^k_{sound} &= p_{max} e^{-\sigma\left(\frac{|j-c_j|}{\mathcal{O}_{width}}\right)^2} e^{-\sigma\left(\frac{|i-c_i|}{\mathcal{O}_{height}}\right)^2} \\ c_j &= -\text{focal} \cdot \frac{\sin(\theta)}{\sin(\frac{1}{2}\pi)} - \theta + \frac{1}{2} I_{width} \\ c_i &= \frac{1}{2} I_{height} \\ \text{focal} &= \frac{I_{width}}{2 \tan\left(\frac{1}{2}\theta_{\text{fov}}\right)} \end{aligned} \tag{5}$$

where p_{max} and σ are the maximum value and standard deviation parameters.

Scene dynamics (dyn). We extract scene dynamics from visual input (RGB image) using the magnitude of vectors composing optical flow. This feature will contribute to increase the saliency of moving proto-objects in the scene. Given a grey scale image I_{of} that stores the magnitudes of the output of a Bayesian optical flow algorithm [23], we perform dilation[3] and then we compute the saliency value for each proto-object,

$$I_{of} = I_{of} \oplus \mathbf{1}_{3\times 3}$$

$$\zeta^k_{dyn} = \begin{cases} 1 & \text{if } \frac{\sum_{i\in\mathcal{O}^k} I_{of}(i) > \epsilon}{\#i\in\mathcal{O}^k | I_{of} > \epsilon} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Colour bias (C_{bias}). This is an interesting feature that can be tuned to produce a biasing behaviour depending on the task being performed. In the case of including a top-down modulation this can be used to improve visual search. In order to model human phylogenetic bias towards red [10], $\tilde{C}_{ref}$ is initially set to pure red.

$$\zeta^k_{C_{bias}} = dist_{HSI}(\tilde{C}_{ref}, \mathcal{O}_{HSI}) \tag{7}$$

Afterwards, the conspicuity maps $\overline{\zeta}_i$, which represent the saliency value of each pixel (i, j) for a specific feature, are then obtained by weighting the saliency of each proto-object with a Gaussian function g with its centre of mass as the mean (μ).

$$\overline{\zeta}_i = \sum_{k\in\mathcal{O}} \zeta^k_i g(\mathcal{O}^k) \tag{8}$$

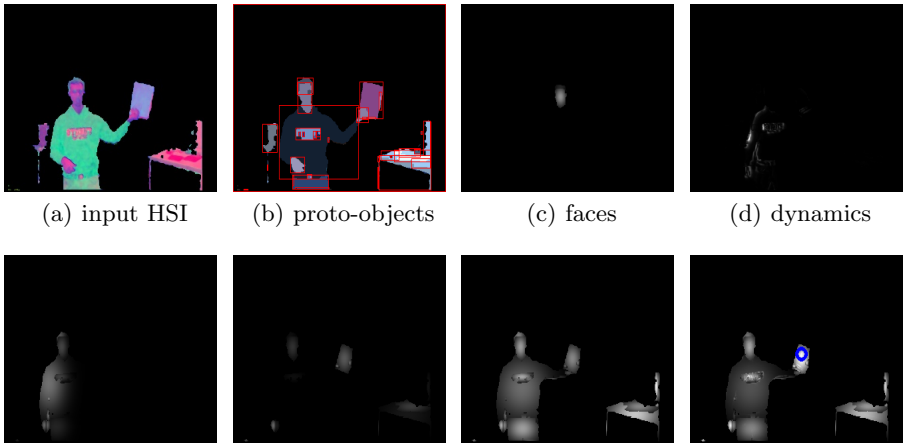
$$g(\mathcal{O}^k) = e^{-\sigma\left(\frac{|j-\mu_j|}{\mathcal{O}^k_{width}}\right)^2} e^{-\sigma\left(\frac{|i-\mu_i|}{\mathcal{O}^k_{height}}\right)^2}$$

Finally, the final 2D saliency map is computed by the weighted summation of the conspicuity maps:

$$\mathcal{S} = \sum_i w_i \overline{\zeta}_i \tag{9}$$

An example of the different conspicuity maps and the final output is depicted in Figure 2, where some interesting outputs are outlined. The input image is transformed to HSI colour space to reduce sensitivity to changes in light conditions. Although the face is composed by two proto-objects (Fig. 2(b)), the face conspicuity map (Fig. 2(c)) shows an unique region that perfectly segments the face. The dynamics (Fig. 2(d)) in this example is attributing too much importance to small proto-objects due to the slight movement of the individual. The auditory conspicuity map, in Fig. 2(e), makes the body of the interlocutor stand out. The colour bias can be seen to clearly emphasise the skin and the red notebook.

[3] Optical flow algorithms detect local motion in object edges, while what we want is to know which proto-object is currently moving.



(a) input HSI (b) proto-objects (c) faces (d) dynamics

(e) auditory (f) colour bias (g) intensity (h) 2D saliency

Fig. 2. 2D saliency computation. (a) image in HSI colour space; (b) proto-object ($\mathcal{O}$) segmentation is shown – each $\mathcal{O}$ is represented using its average colour (bounding boxes are also plotted for better visualisation); (c, d, e, f, g) show different features associated with the $\mathcal{O}$ (colour contrast, which is not shown, is also used); (h) shows the final 2D saliency map and the $\mathcal{O}$ with maximum saliency (blue circle).

3.2 3D saliency

We propose a hierarchical Bayesian approach for constructing the 3D saliency $P(S_c^t)$ map, depicted in Fig. 3. The BVM algorithm, which provides a log-spherical occupancy probability grid $P(O_c)$, is combined with a Bayesian saliency model to obtain a 3D representation that symbolises the important places to be attended. The advantages of using this Bayesian approach are: we can handle the uncertainty of the input cues, filter the noise, perform local prediction and have persistent knowledge that changes with time. This cortically-inspired log-spherical representation of the data allows the system to have a helpful short-memory for top-down processes and for promoting better attention interaction between robots and humans.

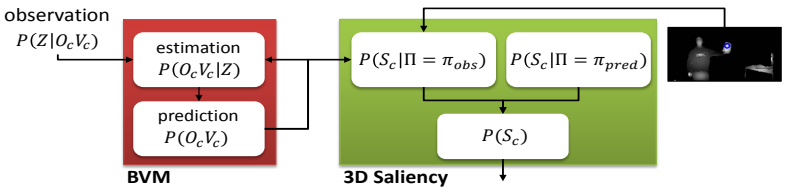


Fig. 3. Hierarchical Bayesian model for 3D saliency. The Bayesian filter that process the BVM [5] is used, combined with the 2D saliency S, to compute the 3D saliency.

To obtain the 3D saliency map we perform the following Bayesian inference:

$$P(S_c^t) = \sum_{\Pi} P(\Pi) P(S_c^t|\Pi) \tag{10}$$

$$P(S_c^t|\Pi = \pi_{obs}) = \sum_{O_c} P(O_c) \max_{i \in c} P(\mathcal{S}|O_c) \tag{11}$$

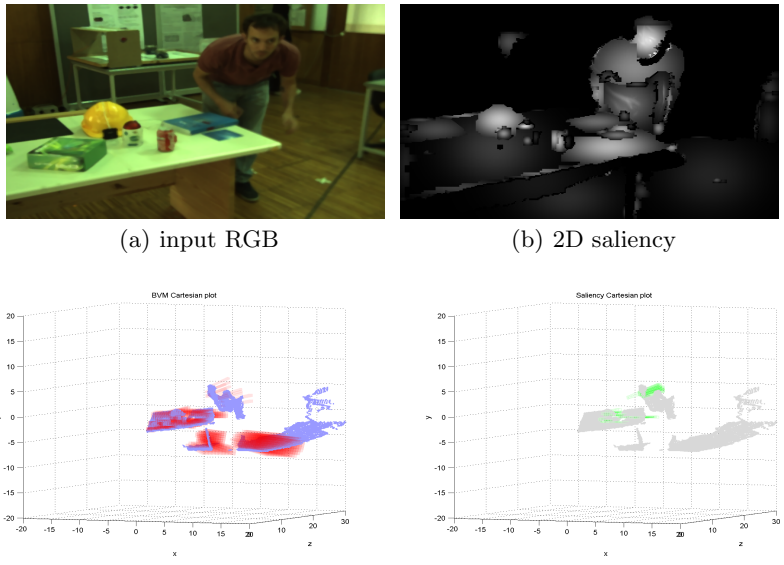
$$P(S_c^t|\Pi = \pi_{pred}) = \sum_{V_c} \frac{\beta_c(S_c^t, [V_c = v_k])}{\sum_{i \in (0,1)} \beta_c([S_c^t = i], [V_c = v_k])} \tag{12}$$

$$\begin{aligned} \beta_c([S_c^t = 1], [V_c = v_k]) &= (1-\alpha) P(A_c|[V_c = v_k]) P(V_c|A_c) P([S_c = 1]|A_c) \\ &+ \alpha P(A_c|[V_c = v_k]) P(V_c|A_c) P([S_c = 0]|A_c) \end{aligned} \tag{13}$$

$$\begin{aligned} \beta_c([S_c^t = 0], [V_c = v_k]) &= \alpha P(A_c|[V_c = v_k]) P(V_c|A_c) P([S_c = 1]|A_c) \\ &+ (1-\alpha) P(A_c|[V_c = v_k]) P(V_c|A_c) P([S_c = 0]|A_c) \end{aligned} \tag{14}$$

where $\alpha \in (0, 1)$ defines the probability of non-constant velocity.

4 Results and Implementation

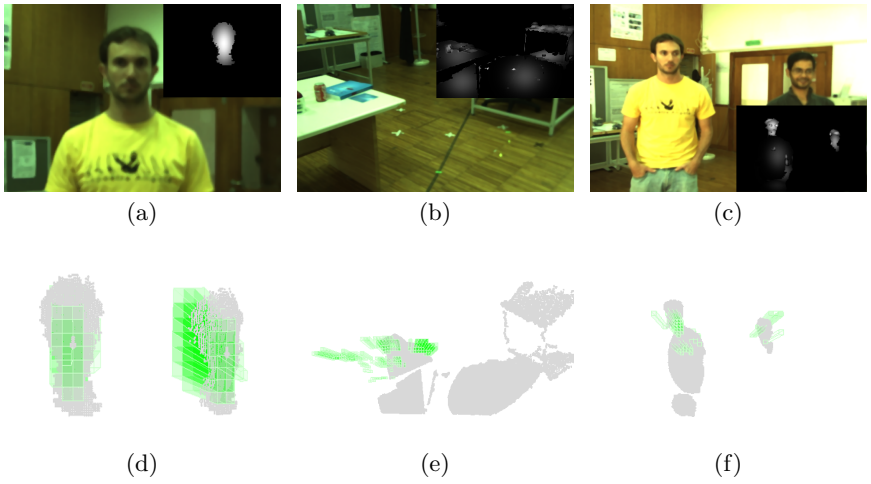


(a) input RGB (b) 2D saliency

(c) BVM (d) 3D saliency

Fig. 4. 3D saliency computation. (a) smoothed input image; (b) 2D saliency; (c) BVM (red 3D trapezoids represent locations with high occupancy probability and colour intensity is proportional to that probability); and (d) final 3D saliency map (green 3D trapezoids represent salient locations and colour intensity describes their relative importance). In (c, d), the point cloud is overlaid by the log-spherical grid maps.

We experimentally analyse the advantages of using the proposed approach for attention systems. The transformation from the original input image to the 3D saliency representation is described in Fig. 4. The input image in RGB is smoothed to avoid salt and pepper noise due to indoor light. The method presented in section 3.1 is applied to obtain Fig. 4(b). Note that the face and the yellow helmet are the most salient proto-objects and their segmentation against the rest of the regions of the scene is clear-cut. Figure 4(b) shows an instant of the BVM running using as the input the point cloud of a depth sensor. The intense red describes the regions that are certainly occupied such as the floor, the table, the objects and the table leg. Conversely, in the final saliency 3D map on Fig. 4(d), only salient regions with high probability of occupancy appear. Bright green corresponds to the individual face, and also objects on the table, such as the yellow helmet.



(a) (b) (c)

(d) (e) (f)

Fig. 5. Properties of the proposed 3D saliency. The first row shows 2D saliency, and the second row the associated 3D representation. (a,d) shows that the maximum salience is in the centre of mass of the proto-object, in this case a face, promoting the correct fixations; (b,e) describes how the short-memory maintains important locations of the scene; and (c,f) shows how the auditory source increases the saliency of one of the interlocutors.

Figure 5 shows some of the properties of the proposed approach. Figure 5(d) shows the proto-object describing a face and the associated saliency in log-spherical representation, in which the centre of the face exhibits the highest value. In Fig. 5(e) panoramic memory persistence maintains salient places outside the current field of view of the robot. It shows the 3D map after rotating the robot head in θ to the right. In this figure, the few salient cells with no point cloud associated on the left side of the figure correspond to previous estimations of the filter for measurements of the table and its objects. Figure 5(f) shows

how visual and auditory cues can help disambiguate behavioural relevance, and important feature for HRI. For instance, two people are in front of the robot and one of them is talking. The saliency associated with the face of the current interlocutor is higher due to the auditory signal.

The proposed approach has been tested in a robotic platform with an active head. We have used the Robotic Operating System (ROS) to integrate all the functionalities. Proto-object segmentation is performed using as inputs images taken using the RGB camera of a Microsoft Kinect, for which the colour space is transformed into HSI to improve robustness against lighting conditions changes. The algorithm employed is an improved implementation of the BIP [15] (e.g., some of the functions have been parallelised). The face feature is extracted using OpenCV Haar classifiers implemented in GPUs, detecting faces in both profile and frontal poses. The dynamics of the scene is estimated using a Bayesian optical flow algorithm [23]. Auditory segmentation provides the azimuth angle of a loud sound source, and is implemented using Hark [16]. Figure 6 shows the computational performance and module dependency details of the system. Note that the implementation deals with the asynchronous nature of the different signals. When testing the initial segmentation algorithms such as the BIP, we obtain faster execution times that running the system all together, for which we have 8 fps for the 2D saliency computation and 3 fps for the 3D representation. The same methodology proposed in [6] can be adapted easily to parallelise the approach into GPUs reaching high performance in real-time. The bottleneck in the system is the face detection, which will be overcome in future work by employing tracking techniques or using the predictive property of the volumetric map.

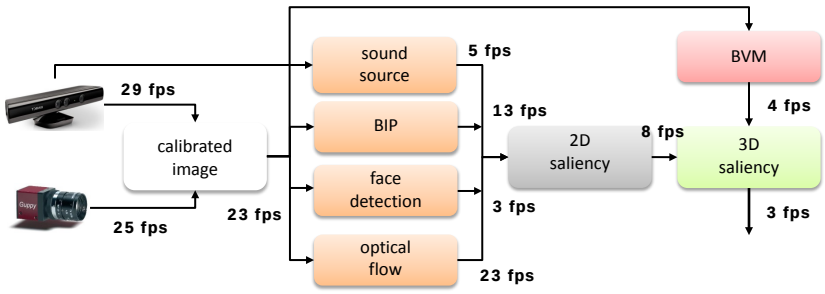


Fig. 6. Computational time analysis of the different stages.

5 Conclusions

We have proposed a 3D saliency computational model that takes into account both visual and auditory cues, and can be used as a short memory in artificial

attention systems. This work presents, as a proof-of-concept, a way to encode regions with behavioural relevance while being robust to noise, ambiguity and uncertainty. The results show that it facilitates the abstraction from saliency to potentially attended objects, by explicitly associating the spatial probability of being salient with a proto-object. It has been also shown that there is persistence of salient regions through time even if the robot is fixating other locations and that the fixations are drawn to the middle of the proto-object. An additional advantage of the system is that faces and auditory sources are useful information for HRI, as shown in the experiments.

As future work, to fully test the potential of the proposed approach, we will integrate 3D saliency into a complete artificial attention system and then perform a systematic evaluation [11]. We want to analyse if attention systems can improve robot adaptivity in open-ended HRI scenarios. For that purpose, we are already extending the system to introduce auditory sources directly into the BVM, as in previous work, to model probable occupied regions outside the field of view, and developing fast algorithms implementation (e.g., GPUs) to increase the system performance up to 15 fps. Moreover, we are investigating ways of modelling the inhibition of return by measuring the entropy associated and how to compute the input signals novelty by means of the 3D saliency representation. Furthermore, we are designing a biologically inspired attention system with top-down modulation to complete the overall framework.

Acknowledgments. 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/EEIAUT/3010/2012]. The authors would like to thank Beatriz Oliveira and Rohit Chandra for their help in the experimental sessions.

References

1. Bruce, N.D., Tsotsos, J.K.: Saliency, attention, and visual search: An information theoretic approach. Journal of vision 9(3), 5 (2009)
2. Corbetta, M., Patel, G., Shulman, G.L.: The reorienting system of the human brain: from environment to theory of mind. Neuron 58(3), 306–324 (2008)
3. Ferreira, J.F., Castelo-Branco, M., Dias, J.: A hierarchical Bayesian framework for multimodal active perception. Adaptive Behavior 20(3), 172–190 (June 2012)
4. Ferreira, J.F., Dias, J.: Attentional Mechanisms for Socially Interactive Robots – A Survey. IEEE Transactions on Autonomous Mental Development 6(2), 110–125 (2014)
5. Ferreira, J.F., Lobo, J., Bessire, P., Castelo-Branco, M., Dias, J.: A Bayesian Framework for Active Artificial Perception. IEEE Transactions on Cybernetics (Systems Man and Cybernetics, part B) 43(2), 699–711 (April 2013)
6. Ferreira, J.F., Lobo, J., Dias, J.: Bayesian real-time perception algorithms on GPU — Real-time implementation of Bayesian models for multimodal perception using CUDA. Journal of Real-Time Image Processing 6(3), 171–186 (September 2011)
7. Frintrop, S., Rome, E., Christensen, H.I.: Computational visual attention systems and their cognitive foundations: A survey. ACM Trans. Appl. Percept. 7(1), 6:1–6:39 (Jan 2010)

8. Garcia-Diaz, A., Leborán, V., Fdez-Vidal, X.R., Pardo, X.M.: On the relationship between optical variability, visual saliency, and eye fixations: A computational approach. Journal of vision 12(6), 17 (2012)
9. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on pattern analysis and machine intelligence 20(11), 1254–1259 (1998)
10. Kuniecki, M., Pilarczyk, J., Wichary, S.: The color red attracts attention in an emotional context. an erp study. Frontiers in human neuroscience 9 (2015)
11. Lanillos, P., Ferreira, J.F., Dias, J.: 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)
12. Lanillos, P., Gan, S.K., Besada-Portas, E., Pajares, G., Sukkarieh, S.: Multi-uav target search using decentralized gradient-based negotiation with expected observation. Information Sciences 282(0), 92 – 110 (2014)
13. Lewis, M.B., Ellis, H.D.: How we detect a face: A survey of psychological evidence. International Journal of Imaging Systems and Technology 13(1), 3–7 (2003)
14. Marfil, R., Palomino, A.J., Bandera, A.: Combining segmentation and attention: a new foveal attention model. Frontiers in computational neuroscience 8 (2014)
15. Marfil, R., Rodrıguez, J., Bandera, A., Sandoval, F.: Bounded irregular pyramid: a new structure for color image segmentation. Pattern Recognition 37(3), 623–626 (2004)
16. Nakadai, K., Takahashi, T., Okuno, H.G., Nakajima, H., Hasegawa, Y., Tsujino, H.: Design and implementation of robot audition system'hark'open source software for listening to three simultaneous speakers. Advanced Robotics 24(5-6), 739–761 (2010)
17. Palomino, A.J., Marfil, R., Bandera, J.P., Bandera, A.: 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)
18. Quak, M., London, R.E., Talsma, D.: A multisensory perspective of working memory. Frontiers in human neuroscience 9 (2015)
19. Rao, S.: Visual routines and attention. Ph.D. thesis, Massachusetts Institute of Technology (2002)
20. Scassellati, B.: Theory of mind for a humanoid robot. In: Humanoid Robotics, 2000 IEEE/RSJ International Conference on (2000)
21. Schauerte, B., Kuhn, B., Kroschel, K., Stiefelhagen, R.: Multimodal saliency-based attention for object-based scene analysis. In: Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on. pp. 1173–1179. IEEE (2011)
22. Schillaci, G., Bodiroža, S., Hafner, V.V.: Evaluating the effect of saliency detection and attention manipulation in human-robot interaction. International Journal of Social Robotics 5(1), 139–152 (2013)
23. Simoncelli, E.P.: Bayesian Multi-Scale Differential Optical Flow. Academic Press (1999)
24. Vossel, S., Geng, J.J., Fink, G.R.: Dorsal and ventral attention systems distinct neural circuits but collaborative roles. The Neuroscientist 20(2), 150–159 (2014)
25. Yanulevskaya, V., Uijlings, J., Geusebroek, J.M., Sebe, N., Smeulders, A.: A proto-object-based computational model for visual saliency. Journal of Vision 13(13), 27 (Nov 2013)