On the advantages of foveal mechanisms for active stereo systems in visual search tasks

Rui Pimentel de Figueiredo[1] · **Alexandre Bernardino**[1] · **José Santos-Victor**[1] · **Helder Araújo**[1]

Received: 15 February 2016 / Accepted: 11 January 2017 / Published online: 4 February 2017

Abstract In this work we study how information provided by foveated images sampled according to the log-polar transformation can be integrated over time in order to build accurate world representations and accomplish visual search tasks in an efficient manner. We focus on a specific visual information modality depth and on how to store it in a flexible memory structure. We propose a probabilistic observational model for a stereo system that relies on the Unscented Transform in order to propagate uncertainty in stereo matching, due to spatial quantization in the retina, to the 3D Cartesian domain. Probabilistic depth measurements are integrated in a novel Sensory Ego-Sphere whose topology can be biased with foveal-like distributions, according to the autonomous agent short-term tasks and goals. Furthermore, we investigate an Upper Confidence Bound algorithm for the task of simultaneously finding the closest object to the observer (visual search) and learning the surrounding environment 3D map (mapping). The performance of task execution is assessed both with a foveated log-polar sensor and a classical uniform one. The advantage of foveal vision and custom ego-sphere representations are illustrated in a series of experiments with a realistic simulator.

Keywords Stereoscopic vision · Foveal vision · Active vision · Sensory ego-sphere

1 Introduction

The spatial organization of the photo-receptors in the human retina is not uniform. Visual acuity is highest at the fovea and declines monotonically to the periphery with increasing eccentricity. This space-variant resolution perception phenomenon—named foveation—is a hardwired mechanism and a natural way of reducing the amount of information streamed to the brain, in order to cope with neuronal transmission bandwidth limitations and the brain machinery processing capacity. However this compression phenomenon introduces a space variant uncertainty either in low-level visual processes as 3D reconstruction, or in higher-level object classification and recognition tasks. In order to efficiently explore and understand the surrounding environment (Posner 2012), humans have developed a set of visual attention and oculo-motor mechanisms that allow them to actively direct the eyes towards different locations in the surrounding environment, thus cleverly compensating for the aforementioned sensory and computational limitations.

Likewise, robots deployed in everyday environments, are faced with increasingly complex scenarios where objects are arranged in many possible different spatial configurations. Moreover, the problem of deciding which regions in the visual field are to be attended during the visual search task is computational demanding. Therefore, like biological systems, robots should be endowed with a set of mechanisms that allow them to search for objects of interest and build detailed maps of the scene, while avoiding the potential computational overload of processing irrelevant sensory stimuli. Under the assumption that biological systems perform quasi-optimally in their environment due to multiple generations of genetic improvement, researchers have been developing biologically inspired systems (Colombo

This is one of several papers published in *Autonomous Robots* comprising the Special Issue on Active Perception.

✉ Rui Pimentel de Figueiredo
ruifigueiredo@isr.ist.utl.pt

[1] Institute for Systems and Robotics (ISR/IST), LARSyS, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal

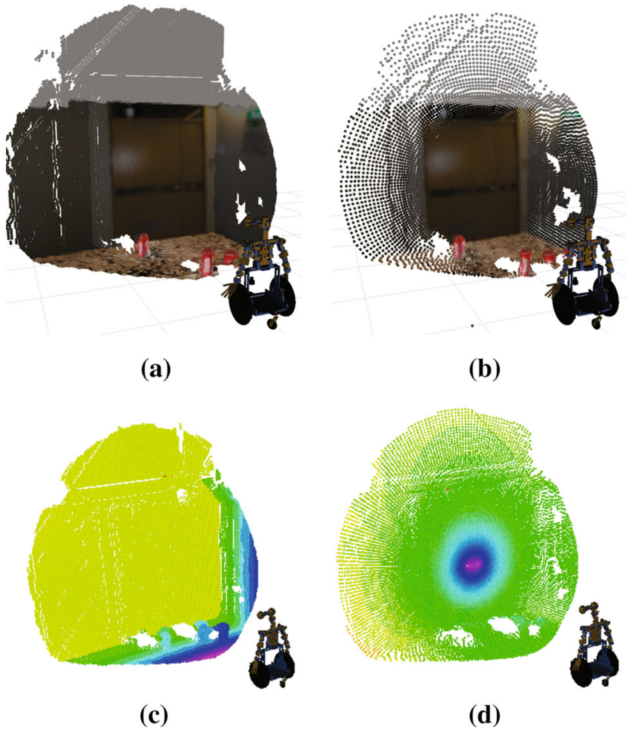

Fig. 1 A snapshot of the RGB-D point clouds (*top row*) and associated probabilistic measures (*bottom row*) obtained with the proposed Cartesian (*left column*) and foveal (*right column*) stereo sensor models. *Blue* and *purple colors* correspond to higher precision measurements. **a** RGB-D, **b** RGB-D, **c** uncertainty and **d** uncertainty (Color figure online)

et al. 1996) equipped with smart multi-resolution sensor topologies and provided with computational models of selective attention (Vijayakumar et al. 2001). These implementations not only mimic the mechanisms observed in humans but also lead to more effective behaviours with constrained resources (computational, energetic), which is one of the main goals to pursue of today's cognitive robotics.

In this work we propose a probabilistic selective attentional framework for artificial systems provided with binocular foveal vision. Our framework relies on visual information and associated confidence measures (see Fig. 1) that are used to autonomously drive the agent's gaze direction during search tasks. Our contributions are the following. First, we model the stereo reconstruction uncertainty that arises as a result of spatial quantization phenomena inherent in the retina. Our approach considers Gaussian Receptive Fields[1] (RFs) distributed in space following two different tessellations: (i) a classical uniform (Cartesian) arrangement and (ii) a log-polar one that mimics the human retina. The RFs in the latter present a space-variant spatial distribution and support radius (Pamplona and Bernardino 2009). The Unscented Transform (UT) (Julier and Uhlmann 2004) is used to propagate belief from the 2D retina domain to 3D via stereo reconstruction. When compared with previous approaches that also assume Gaussian quantization noise and that rely on first order linearizations to approximate the non-linear transformations involved in 3D reconstruction (Kriegman et al. 1989), our method based on UT is more precise and hence improves 3D estimation quality. Second, the probabilistic sensory measurements are integrated in a novel versatile randomized Sensory Ego-Sphere (SES) whose topology can be biased according to the autonomous agent short-term tasks and goals. The proposed SES, helps achieving the task, by allocating the limited resources to important surrounding regions according to the task. Finally, a decision-making process, framed within a multi-armed bandit setting (Auer et al. 2002), acts as a mediating cognitive attentional process that seeks to maximize expected task-related rewards. The proposed decision algorithm relies on statistical measures to decide where to look next by selecting the most promising regions to attend. We investigate a simple Upper Confidence Bound (UCB) algorithm (Agrawal 1995) for the task of finding the closest object to the observer. The UCB algorithm controls the exploration–exploitation trade-off typical of decision under uncertainty algorithms: to accomplish the task it is necessary to explore the world, but too much exploration will delay the task execution.

The remainder of this paper is organized as follows. In Sect. 2 we conduct a brief overview of the attentional frameworks available in the literature with a strong emphasis on probabilistic-based methodologies. In Sect. 4, we outline the proposed sensor observation model and the uncertainty propagation model from the retinal domain to 3D. In Sect. 5, we introduce a novel biologically inspired short-term memory structure which is egocentric, compact, and convenient for fast and efficient information update and retrieval. In Sect. 6, we endow our system with a decision-making process that actively drives the agent's gaze direction, through sequential saccadic eye movements. Finally, in Sect. 7, we experimentally validate our model and compare a conventional Cartesian camera against a space-variant vision system. The obtained results demonstrate that a wider field of view at the cost of less peripheral resolution is advantageous in visual-search tasks. We show that with our methodologies different gaze patterns emerge depending on the sensor characteristics and decisions on confidence bounds. Furthermore, we demonstrate that spatial memory biases, reflecting prior knowledge about the world structure and the task at hand, allow large performance improvements in visual search tasks.

[1] *Receptive fields* are the fundamental visual processing units. Each corresponds to a specific region in the retina (image) and is represented by the average value of the photo-receptors (pixels) within it (e.g. average color). For more details, we refer the interested reader to Edelman (1995).

2 Related work

Probabilistic based active vision requires not only the characterization of the sensory-motor uncertainties, but also the definition of memory structures that facilitate continuous recall and temporal fusion of probabilistic sensory data. Therefore, we organize the present section in two distinct parts. At first we overview the state-of-the-art in active vision with an emphasis on probabilistic models of overt attention. Afterward, we analyse the memory data structures proposed in the literature suitable for applications related to attention.

2.1 Overt attention and active vision

The first studies on visual attention date back to the mid nineteenth century, pioneered by von Helmholtz and König (1896) and motivated by the willingness to understand how humans attended stimuli at the periphery of the visual field. Nowadays, the literature on selective attention is vast, and covers a wide range of scientific fields, including cognitive neuroscience (Carrasco 2011) and computer science (Borji and Itti 2013), playing an important role in computer vision and robotics applications (Begum and Karray 2011). In this work we focus on a particular aspect of attention: active sensing which is tightly coupled to the concept of overt attention. The goal of active vision systems is to direct the eyes towards locations such that:

- the information about the surrounding environment is increased over time (exploration);
- the desired region is centered in the eyes and thus observed by the retinal zone of maximum visual acuity (exploitation).

Early work on computational modeling of overt attention (Koch and Ullman 1987) suggests that saccadic eye movements are attracted in a bottom-up manner to salient stimuli, i.e. to areas of the the visual field that differ from the neighborhood in different feature modalities and spatial scales, in a center surround fashion. Arguably, the most influential work on saliency modeling was proposed in Itti et al. (1998) where the authors introduced a unique spatial saliency map which encoded the prominent locations over the entire visual scene, to be attended after and analyzed in detail in order of decreasing conspicuity. A major drawback of the saliency models is that the classical sequencing mechanism (inhibition of return) only considers spatial information (location) and not of the information provided by the stimuli. Therefore, they do not not leverage information from previous steps to improve information on unexplored areas. Ongoing sensory data gathering while attending salient regions should continuously affect reasoning and influence decisions of where to look next. This decision process requires not only a continuous evaluation of new stimuli arising from previously not visible regions, as well as reassessing previously attended locations as a result of the newly acquired information.

More sophisticated models are framed within probabilistic paradigms that account both for sensori-motor uncertainties as well as the world intrinsic stochasticity and unpredictability. A common idea behind these models is that statistical objectives are the fundamental driving elements behind visual attention. From a Bayesian standpoint, attention seeks to actively infer the future actions that maximize the expected information gain given the spatio-temporal context. Therefore informational gain is itself the inner goal behind attention (Friston et al. 2012).

The probabilistic-based saliency model proposed in Itti and Baldi (2006) suggests that surprising events or stimuli attract attention. The Kullback–Leibler (KL) divergence between prior and posterior beliefs is by convention used as a measure of surprise. However, surprise models are purely exogenous by nature since they react to observed stimuli. Active vision models based on optimal stochastic control principles pose the action selection problem within Bayes risk minimization framework, and differ on the chosen policies. On one hand, infomax algorithms (Butko and Movellan 2010) seek to maximize the expected accumulated future informational gain in fixed time-horizon. On the other hand greedy MAP policies consider only a one-step look ahead time window (Najemnik and Geisler 2005) and self-knowledge about the retinal acuity map to decide the best location to attend. A recent work on active sensing accounted also for behavioral costs (Ahmad and Yu 2013), such as the energy and temporal costs incurred in choosing a given motor action.

Despite the demonstrated applicability of the previously mentioned approaches on target search tasks in monocular images, there are no works studying depth cues inferred by stereo vision, and the influence of foveal vision in the search strategies on binocular setups. The stereo reconstruction problem using foveated images has been addressed in the literature, namely in Bernardino and Santos-Victor (2002), where the authors have shown that it is possible to compute dense disparity maps from log-polar images. Nevertheless, with foveal images, stereo matching accuracy degrades in the image periphery. This motivates the need for modeling depth uncertainty in stereo reconstruction, due to space-variant discretization in foveated images and use this uncertainty to decide where to look next. In this paper we analyze the ability of active foveal stereo systems to accurately map the environment and efficiently execute visual search tasks. Some visual tasks are more naturally represented in 3D, for instance the search for nearby objects, as illustrated in this work. Therefore, the main contribution of the paper is the formulation of visual search tasks in 3D and a the development of novel methods for uncertainty propagation and spatial representations required for this purpose. We show that adequate retinal topologies and 3D spatial representation play a role in the speed of execution and accuracy of localization of targets in 3D search tasks.

2.2 Spatial memory data structures

Memory plays a key role and is a core component of any cognitive architecture. In real-time decision-making problems involving perception, autonomous agents rely on memory structures to store and query continuously obtained information in a robust and efficient manner.

The Sensory Ego-Sphere (SES) is an egocentric, short-term memory structure (Peters Ii et al. 2009) that is convenient for sensory data fusion and that has been extensively used in robotics applications involving attention (Ruesch et al. 2008; Fleming et al. 2006). In the attention domain and from a practical point of view, egocentric spherical representations offer several advantages in applications involving humanoid robots when compared to typical Cartesian representations such as regular occupancy grids, point clouds or octrees (Hornung et al. 2013). Spherical representations based on egocentric polar coordinate systems, are typically more compact (low memory requirements) and avoid the requirement of computationally expensive ray-casting techniques when dealing with visibility issues.

Different representations and data structures for the SES have been proposed in the literature (see Fig. 4). Typically, 2D array type structures based on spherical coordinate systems are used to represent the spherical surface (Ruesch et al. 2008). These can be accessed in $\mathcal{O}(1)$ time and thus are appropriate for real-time applications. Yet, they are non-isotropic and therefore data is not stored uniformly over the surface (i.e. the resolution is higher near the poles). On the other hand, the geodesic dome type data structure (Peters Ii et al. 2009) is isotropic and therefore can better approximate 3D shape. However, indexing becomes less trivial and less efficient due to its non-regular topology. To tackle this issue Hirose et al. (2002) proposed a hierarchical geodesic structure that can significantly speed-up access times. In another work Ferreira et al. (2008) proposed an egocentric log-spherical grid named Bayesian Volumetric map occupancy spherical grid, that was proven suitable for probabilistic multi-modal sensor fusion.

Nevertheless, none of the previous mentioned structures are easily reconfigurable for the implementation of task-dependent cognitive biases that either enhance or impair the storage and recall of information in short-term spatial memory (Crawford et al. 2014). This fact motivates the need for more sophisticated, task-biased, versatile memory structures. In Sect. 5 we propose a novel memory representation that tackles this problem. Typical tessellations of the sphere include quasi-uniform icosahedral tessellations, less uniform spherical polyhedra or non-uniform latitude/longitude grids. All these forms are highly regular and structured, which limits their flexibility to implement arbitrary shapes. The method proposed in this paper is based on projecting in the sphere randomly generated points according to a mixture of 3D gaussian distributed points with an arbitrary number of components, focal points (means) and dispersions (covariances). This generates an irregular grid but we can define more freely areas on the sphere with varying degrees of density and dispersion. Our sampling scheme is easy to implement and allows for the fast creation of task-biased sensory ego-spheres. As opposed to the previously proposed deterministic counterparts, our SES relies on an easy to implement random sampling scheme that allows for the fast creation and real-time access of arbitrary reconfigurable topologies.

3 Problem statement and system overview

In the proposed problem, the observer's goal is to select the oculomotor actions that maximize task related rewards. On one hand we rely on a recursive Bayesian filter that sequentially accumulates sensory inputs and extracts valuable information about the agent and the environment state, given noisy observations. On the other hand, a decision-making algorithm predicts the best future locations to gather information, according to some statistical or behavioral criteria.

The environment structure, i.e. 3D map, is a projection of the world structure $W \subset \mathbb{R}^3$ in the agent's egocentric reference frame $\mathcal{E}$, internally represented by a discrete set of points, each associated to a specific observation direction (see Sect. 4). Let us denote the set of environment sample points by

$$X_t = \left\{ \mathbf{x}_t^i \in \mathbb{R}^3, i = 1, \ldots, N_x \right\} \tag{1}$$

where N_x is the total number of considered observation directions. These points are modeled as Gaussian random variables, initialized with mean and covariance selected according to a priori knowledge about the type of environment in which the robot operates. The egocentric reference frame $\mathcal{E}$ is head-centered, has three translational degrees of freedom and fixed orientation with respect to the world frame of reference (see Fig. 2).

In order to execute visual search tasks, the proposed cognitive architecture is equipped with two sensory-motor modalities:

- *proprioception* provided by odometric and oculocephalic joint encoders;
- *stereo vision* provided by a stereo camera system.

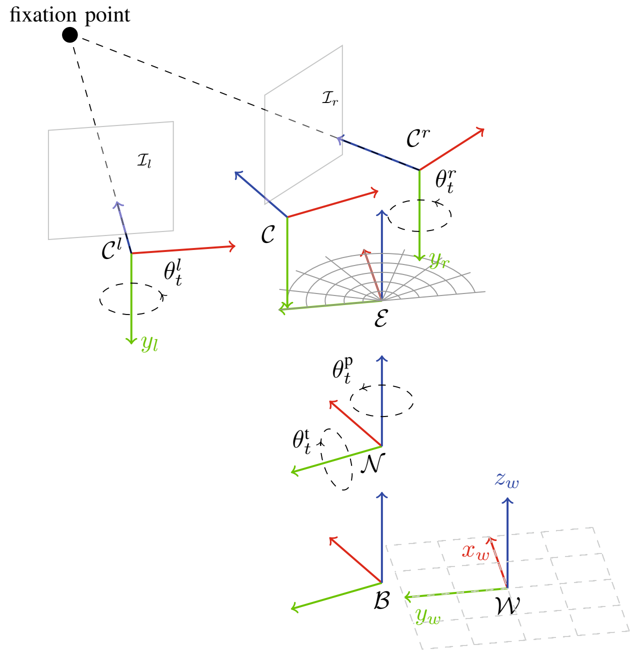

Fig. 2 The various coordinate systems used by our system (best seen in *color*): the inertial world frame ($\mathcal{W}$) in which the environment is represented; the base frame ($\mathcal{B}$) which is rigidly attached to the mobile robot base, and permits determining the robot pose in the world, given the odometric readings; the neck frame ($\mathcal{N}$) which allows representing pan and tilt cephalic movements; the egocentric frame ($\mathcal{E}$) in which spatial memory is embodied and sensor fusion is performed; the cyclopean frame ($\mathcal{C}$) in which stereo observations are represented; the convergent, non-parallel pair of cameras frames ($\mathcal{C}^l$, $\mathcal{C}^r$), in which monocular images are obtained (Color figure online)

The observer is allowed to change its state, i.e. the observation view point, through base and oculocephalic movements (please see Fig. 1). At each time instant, the proprioceptive modality reports the robot base location and its internal kinematic state. More specifically, the robot position and orientation $P \in \mathbb{R}^6$ in the inertial frame of reference $\mathcal{W}$, the agent's eyes horizontal vergence ($\theta_t^{\mathrm{v}} \in \mathbb{R}$) and the head pan and tilt joint angles ($\theta_t^{\mathrm{p}}, \theta_t^{\mathrm{t}} \in \mathbb{R}$). Let us denote the joint set of odometric and oculocephalic measured/controlled joint positions by

$$U_t = \left\{P_t, \theta_t^{\mathrm{v}}, \theta_t^{\mathrm{p}}, \theta_t^{\mathrm{t}}\right\} \tag{2}$$

We assume that the proprioceptive modality provides noise-free observations. In other words, we consider that the measurement errors are negligible with respect to the visual sensor errors and therefore that the robot location and kinematics, and thus, the transformations between the various reference frames involved in our system (see Fig. 2), can be deterministically determined from U_t. Furthermore, we assume that the environment W is static for the duration of the search task and is not affected by the robot motor actions U_t (the base location and the posture of the robot's head). The preceding assumptions yield the following probabilistic simplification

$$p(X_t|W, U_t) = p\left({}^{\mathcal{E}}\mathbf{R}_{\mathcal{W}} W + {}^{\mathcal{E}}\mathbf{t}_{\mathcal{W}}\right) = p(X_t) \tag{3}$$

where ${}^{\mathcal{E}}\mathbf{R}_{\mathcal{W}} \in \mathbb{R}^{3\times3}$ and ${}^{\mathcal{E}}\mathbf{t}_{\mathcal{W}} \in \mathbb{R}^{3\times1}$ are an orthogonal rotation matrix and a translation vector, respectively, obtained by combining deterministic proprioceptive joint angle measurements with known forward kinematics.

The stereo sensor computes a list of 3D point estimates defined in a cyclopean reference frame $\mathcal{C}$, with origin at the midpoint of the stereo baseline, from noisy point correspondences observed in the left and right retinal domain. Let us denote the set of 3D points by

$$Z_t = \left\{\mathbf{z}_t^o \in \mathbb{R}^3,\ o = 1, \ldots, N_{v,t}\right\} \tag{4}$$

where $N_{v,t}$ is the total number of observed independent and identically distributed (iid) measurements by the stereo sensor at time t. The observation model described in Sect. 5 explains how measurements Z_t are generated according to the environment 3D structure, egocentric projection X_t:

$$Z_t \sim p(Z_t|X_t) \tag{5}$$

4 Stereo sensor model

In stereo vision, a general stereo matching algorithm computes a set of one-to-one point correspondences between two images (Tippetts et al. 2016). However the precision of the measurements is finite and constrained by the fundamental image-sensing units size and spacing. In order to model reconstruction uncertainty due to the limited sensing precision at the retinal level we consider a probabilistic observation model for our stereoscopic sensor (Perrollaz et al. 2010).

4.1 Nonparallel stereo system

Let us suppose that our stereoscopic system is composed by a convergent, non-parallel pair of pinhole cameras $\mathcal{C}^l, \mathcal{C}^r$, allowed to rotate around their y optical-axis by $\theta^l = \frac{\theta^v}{2}$ and $\theta^r = -\frac{\theta^v}{2}$, respectively, and are separated by a fixed baseline b. Furthermore, let us assume that the stereo system is calibrated, and thus, the intrinsic $\mathbf{K}^l$, $\mathbf{K}^r$ and extrinsic $\mathbf{R}_y\left(\theta^v\right)$, $\mathbf{T}\left(b\right)$, camera parameters are always known.

4.2 Gaussian stereoscopic retinal observation model

Let us consider that the cameras image planes $\mathcal{I}^l, \mathcal{I}^r \subset \mathbb{R}^2$ comprise a finite set of RFs denoted by $\mathcal{S}^l, \mathcal{S}^r \subset \mathbb{R}^2$. We

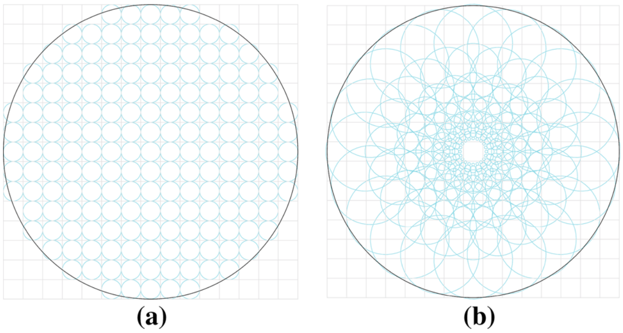

Fig. 3 Gaussian receptive fields with support plotted for 3 standard deviations, **a** Cartesian and **b** log-polar

assume that each RF has a non-uniform stimuli response, modeled by a two dimensional Gaussian profile (Pamplona and Bernardino 2009), with the support regions depicted in Fig. 3. The mean $\boldsymbol{\mu} = \left(\mu_x, \mu_y\right)$ defines the coordinates of the center of the RF in the retinal plane, where response is maximal, and the standard deviation σ represents its support radius.

Thus, observing a correspondence at a given RF pair $\mathbf{s}^i \in \mathcal{S}^l \times \mathcal{S}^r$ follows a conditional Gaussian distribution:

$$\mathbf{s}^i \sim \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{s}^i}, \boldsymbol{\Sigma}_{\mathbf{s}^i}\right) \tag{6}$$

where

$$\boldsymbol{\mu}_{\mathbf{s}^i} = \begin{bmatrix} \mu_x^{l,i} \\ \mu_y^{l,i} \\ \mu_x^{r,i} \\ \mu_y^{r,i} \end{bmatrix}, \quad \Sigma_{\mathbf{s}^i} = \operatorname{diag}\left({\sigma^{l,i}}^2, {\sigma^{l,i}}^2, {\sigma^{r,i}}^2, {\sigma^{r,i}}^2\right) \tag{7}$$

4.3 Stereoscopic reconstruction

Given a point pair correspondence $\mathbf{s}^i$ found in the retinal domain, we determine the corresponding 3D position in the cyclopean reference frame via stereo analysis. Since point pair correspondences are inherently corrupted with precision errors, their projection lines may no longer satisfy the epipolar constraint and therefore not intersect in 3D space. Hence, one should rely on a triangulation method, denoted by τ, in order to compute a 3D Cartesian point estimate $\hat{\mathbf{z}}$ from a point correspondence in image coordinates $\mathbf{s}^i$:

$$\tau : \mathcal{I}^l \times \mathcal{I}^r \longrightarrow \mathbb{R}^3 \tag{8}$$

Due to its simplicity and relatively low computational complexity, we use the mid-point method (for details please refer to Wang and Liu 2007).

4.4 Uncertainty back-propagation via the unscented transform

Since the transformation (8) involved in 3D reconstruction is non-linear, we employ the Unscented transform (Julier and Uhlmann 2004) to compute the propagated mean and covariance up to the third order (by Taylor's expansion). This is achieved by approximating a multivariate Gaussian distributed variable with a set of meaningful and deterministically chosen set of samples (usually named sigma points). For each receptive field pair $\mathbf{s}^i \in \mathcal{S}^l \times \mathcal{S}^r$ we associate a set of sigma points

$$\mathcal{U}^i = \left\{\mathcal{X}^{(i,j)} \in \mathcal{I} \times \mathcal{I}' : j = 0, \ldots, 2N_s\right\} \tag{9}$$

where N_s is the number of sigma points, which are pre-computed according to the following expressions

$$\mathcal{X}^{(i,0)} = \boldsymbol{\mu}_{\mathbf{s}^i} \tag{10}$$

$$\mathcal{X}^{(i,j)} = \boldsymbol{\mu}_{\mathbf{s}^i} + \left(\sqrt{(N_s + \lambda)\Sigma_{\mathbf{s}^i}}\right)_j \quad \text{for} \quad j = 1, \ldots, N_s \tag{11}$$

$$\mathcal{X}^{(i,j)} = \boldsymbol{\mu}_{\mathbf{s}^i} - \left(\sqrt{(N_s + \lambda)\Sigma_{\mathbf{s}^i}}\right)_j \quad \text{for} \quad j = N_s + 1, \ldots, 2N_s \tag{12}$$

where $(\cdot)_j$ denotes the j-th row of a matrix. Furthermore, we consider a set of weights

$$\mathcal{W} = \left\{w_c^{(j)}, w_m^{(j)} \in \mathbb{R} : j = 0, \ldots, 2N_s\right\} \tag{13}$$

which are computed as follows

$$w_m^{(0)} = \frac{\lambda}{L + \lambda} \tag{14}$$

$$w_c^{(0)} = \frac{\lambda}{L + \lambda} + \left(1 - \alpha^2 + \beta\right) \tag{15}$$

$$w_m^{(j)} = w_c^{(j)} = \frac{1}{2\,(L + \lambda)} \quad \text{for} \quad j = 1, \ldots, 2N_s \tag{16}$$

where $\lambda = \alpha^2(L + K) - L$ is a scaling factor, α controls the spread of the sigma points around the mean, K is a secondary scaling parameter, and β is used to incorporate prior knowledge about the distribution of $\mathbf{s}$ (for Gaussian distributions $\beta = 2$ is optimal). Then, for a given point correspondence in retinal domain, we first apply the non-linear transformation τ to the sigma points associated with the corresponding RF pair, $\mathcal{Z}^{(i,j)} = \tau(\mathcal{X}^{(i,j)})$, and then re-estimate the mean and covariance in the 3D domain, according to

$$\hat{\boldsymbol{\mu}}_{\mathbf{z}^i} = \sum_{j=0}^{2N_s} w_m^{(j)} \mathcal{Z}^{(i,j)} \tag{17}$$

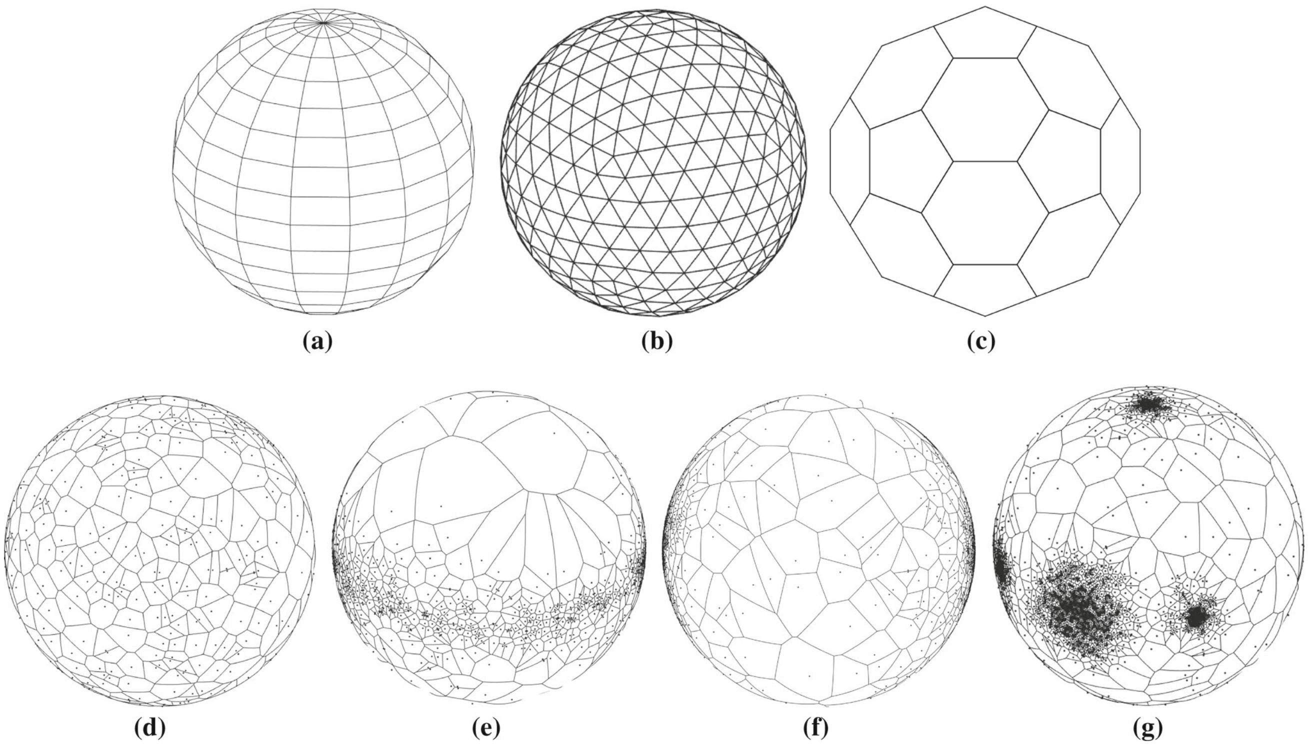

Fig. 4 Different sensory ego-spheres, resulting from different tessellations: *top row* illustrates highly regular, deterministic structures. The *bottom row* depicts our novel randomized structure for different task-dependent biases. **a** 2D array, **b** geodesic dome, **c** truncated icosahedron, **d** unbiased, **e** equator-biased ($M = 1$) and **f** antipodal ($M \geq 1$)

$$\hat{\Sigma}_{\mathbf{z}^i} = \sum_{j=0}^{2N_s} w_c^{(j)} \left(\mathcal{Z}^{(i,j)} - \hat{\boldsymbol{\mu}}_{\mathbf{z}^i}\right)\left(\mathcal{Z}^{(i,j)} - \hat{\boldsymbol{\mu}}_{\mathbf{z}^i}\right)^T \tag{18}$$

The sigma points in retinal domain are computed offline and stored in a linear array in order to speed up on-line uncertainty propagation.

5 Randomized sensory ego-sphere

In the proposed framework the SES plays an intermediate role between the stereo sensor and the decision planning process. Probabilistic data arriving from the sensory stream is continuously fused and integrated over time in the SES, by means of recursive Bayesian filter. At the same time the available information is used to predict and redirect gaze to the best expected location, in the light of new observations.

5.1 Definition

The proposed SES is composed of a set of cells $\mathcal{P}$ lying on a unit sphere, whose center corresponds to a certain absolute orientation , and a map that assigns to each cell the 3D coordinates of the points observed by the robot at that orientation

$$\mathcal{M} : \mathcal{R}^3 \longrightarrow \mathcal{P} \tag{19}$$

The proposed cell grid structure is analogous to a Voronoi diagram defined on a spherical 2-manifold $\mathbb{S}^2$ in 3D space, as depicted in Fig. 4. In practice the proposed SES comprises a set of 3D Cartesian sample points with unit norm and centered in the observer egocentric reference frame $\mathcal{E}$,

$$\mathcal{P} = \left\{\mathbf{p}^i \in \mathbb{R}^3, i, \ldots, N_x : \|\mathbf{p}^i\| = 1\right\} \tag{20}$$

which are i.i.d. and randomly generated from a three dimensional Gaussian Mixture Model (GMM) distribution

$$\mathbf{p}^i = \frac{\mathbf{v}^i}{\|\mathbf{v}^i\|} \quad \text{where} \quad \mathbf{v}^i \sim p\left(\boldsymbol{\theta}\right) = \sum_{m=1}^{M} \phi^m \mathcal{N}\left(\boldsymbol{\mu}_p^m, \boldsymbol{\Sigma}_p^m\right) \tag{21}$$

where M is the number of mixture components and where each $\mathbf{p}^i \in \mathcal{P}$ represents an orientation, allowing for efficient data-alignment with observed 3D points, using inner products (Eq. 23). Each SES cell, represented by $\mathbf{p}^i \in \mathcal{P}$, stores one environment sample point estimate $\mathbf{x}^i \in X$.

The statistics of the Gaussian Mixture Model distribution are chosen according to the observer goals. On one hand, in order to produce uniform and unbiased memory structures, the surface should be sampled from a rotationally symmetric distribution, i.e. from a single Gaussian with zero mean and variance equal in all dimensions (Muller 1959) (Fig. 4d). On the other hand, non-uniform, task-dependent memory biasing can be achieved by manipulating the Gaussian Mixture Model parameters. The proposed randomized representation offers a convenient mechanism for encoding task and world prior knowledge. Memory biasing should lead to more efficient, flexible and adaptable memory allocation and to more effective behaviours during task execution.

Hypothetical topologies that may be suitable for different tasks are depicted in Fig. 4: If for instance the task is to look for people, one should privilege areas at the equator rather than the poles. In this case, varying the Gaussian mean is not sufficient. One could sample from a single-component zero mean GMM with larger variance in the horizontal directions (Fig. 4e). While crossing a street, the observer should prioritize attentional resources to antipodal, lateral regions (Fig. 4f). This can be achieved by sampling from a single-component Gaussian with a larger variance in the lateral component, or from a two-component GMM with opposite lateral means. More complex tasks can benefit from irregular topologies with multiple foci, obtained from GMMs with many components (Fig. 4g).

5.2 Data alignment

For each observed world point estimate provided by our stereo observation model at time t, $\mathbf{z}_t^o$, we need to find the associated memory cell in order to perform probabilistic data fusion. The association process goes as follows. First, the observed random variable is transformed from the cyclopean to the egocentric reference frame, according to the linear transformation $Z' : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ of the form

$$Z' = {}^{\mathcal{E}}\mathbf{R}_{\mathcal{C}} Z + {}^{\mathcal{E}}\mathbf{t}_{\mathcal{C}} \tag{22}$$

where ${}^{\mathcal{E}}\mathbf{R}_{\mathcal{C}} \in \mathbb{R}^{3\times 3}$ is an orthogonal rotation matrix and ${}^{\mathcal{E}}\mathbf{t}_{\mathcal{C}} \in \mathbb{R}^{3\times 1}$ a translation vector, obtained by combining proprioceptive joint angle measurements with known forward kinematics.

Second, for each observation $\mathbf{z}'^{o}_{t}$ we find the associated memory cell c^o, which is the one that minimizes the Euclidean distance, according to the mapping function $\mathcal{M}$, here defined as follows

$$c^o = \mathcal{M}\left(\hat{\boldsymbol{\mu}}_{\mathbf{z}'^{o}_{t}}\right) = \underset{j}{\operatorname{argmin}} < \mathbf{p}^j, \frac{\hat{\boldsymbol{\mu}}_{\mathbf{z}'^{o}_{t}}}{\|\hat{\boldsymbol{\mu}}_{\mathbf{z}'^{o}_{t}}\|} > \tag{23}$$

After finding the associated cell we update its respective estimate according to Eq. (33). Moreover, we assume that the transformed observations are conditionally independent, given X_t, and thus

$$p(Z'_t|X_t) = \prod_{o=1}^{N_v} p\left(\mathbf{z}'^{o}_{t}|\mathbf{x}_t^{c^o}\right) \tag{24}$$

Finally, the resulting probabilistic observation model $p(\mathbf{z}'^{o}_{t}|\mathbf{x}_t^{c^o})$ follows a Gaussian distribution

$$\mathbf{z}'^{o}_{t}|\mathbf{x}_t^{c^o} \sim \mathcal{N}\left(\hat{\boldsymbol{\mu}}_{\mathbf{z}'^{o}_{t}}, \hat{\Sigma}_{\mathbf{z}'^{o}_{t}}\right) \tag{25}$$

with statistics computed as follows

$$\hat{\boldsymbol{\mu}}_{\mathbf{z}'^{o}_{t}} = {}^{\mathcal{E}}\mathbf{R}_{\mathcal{C}} \hat{\boldsymbol{\mu}}_{\mathbf{z}_t^o} + {}^{\mathcal{E}}\mathbf{t}_{\mathcal{C}} \tag{26}$$

$$\hat{\Sigma}_{\mathbf{z}'^{o}_{t}} = {}^{\mathcal{E}}\mathbf{R}_{\mathcal{C}} \hat{\boldsymbol{\Sigma}}_{\mathbf{z}_t^o} \mathbf{R}_{\mathcal{C}}^T \tag{27}$$

5.3 Probabilistic sensor fusion

In the sensor fusion perspective, the goal of the optimal Bayesian estimator is to determine the posterior probability distribution over X, given the accumulated visual sensory observations and the robot proprioceptive state measurements up to time $t \in \mathbb{N}$. Sequential Bayesian filtering allows us to accumulate sensor inputs and update the likelihood of X, at each time instant.

The posterior probability distribution at time t, of the set of internal environment sample points X_t given the current and past visual and proprioceptive observations, is given by

$$p(X_t|Z_{1:t}, U_{1:t}) = p(X_t|Z_t, Z_{1:t-1}, U_{1:t}) \tag{28}$$

Furthermore, since we assume that the proprioceptive measurements are deterministic, then

$$p\left(Z'_t|Z_t, U_t\right) = p\left({}^{\mathcal{E}}\mathbf{R}_{\mathcal{C}} Z_t + {}^{\mathcal{E}}\mathbf{t}_{\mathcal{C}}\right) = p(Z'_t) \tag{29}$$

and Eq. (28) becomes

$$p\left(X_t|Z'_{1:t}\right) = p\left(X_t|Z'_t, Z'_{1:t-1}\right) \tag{30}$$

Since the world is static, at each iteration, the solution to the filter involves only one update step: in the *measurement update* step observations are used to update the current belief by applying the Bayes rule to the right hand side of Eq. (30) and using the observation model (5) we get

$$p\left(X_t|Z'_{1:t}\right) = \eta p\left(Z_t|X, Z'_{1:t-1}\right) p\left(X|Z'_{1:t-1}\right) \tag{31}$$

where η is a normalizing constant. Since the current observations Z'_t are conditionally independent of the past observations $Z'_{t:t-1}$ given the current environment projection in the egocentric frame, X_t, the previous equation becomes

$$p\left(X_t|Z'_{1:t}\right) = \eta p\left(Z'_t|X_t\right) p\left(X_t|Z'_{1:t-1}\right) \tag{32}$$

The a posteriori is independently determined for each cell, according to

$$p\left(\mathbf{x}_t^{c^o}|\mathbf{z'}^{o}_{1:t}\right) = \eta p\left(\mathbf{z'}^{o}_t|\mathbf{x}_t^{c^o}\right) p\left(\mathbf{x}_t^{c^o}|\mathbf{z'}^{o}_{1:t-1}\right) \tag{33}$$

and follows a Gaussian distribution, with statistics given by

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_t^{c^o}} = \left(\hat{\boldsymbol{\Sigma}}^{-1}_{\mathbf{x}^{c^o}_{t-1}} + \hat{\boldsymbol{\Sigma}}^{-1}_{\mathbf{z'}^{o}_t}\right)^{-1} \tag{34}$$

$$\hat{\boldsymbol{\mu}}_{\mathbf{x}_t^{c^o}} = \hat{\boldsymbol{\Sigma}}_{\mathbf{x}_t^{c^o}} \left(\hat{\boldsymbol{\Sigma}}^{-1}_{\mathbf{x}^{c^o}_{t-1}} \hat{\boldsymbol{\mu}}_{\mathbf{x}^{c^o}_{t-1}} + \hat{\boldsymbol{\Sigma}}^{-1}_{\mathbf{z'}^{o}_t} \hat{\boldsymbol{\mu}}_{\mathbf{z'}^{o}_t}\right) \tag{35}$$

Each point estimate is initialized with large mean and covariance to reflect the high uncertainty due to non-existent world prior knowledge.

6 Active sensing: sequential stochastic decision making

In the proposed framework, the decision making is responsible for sensory-motor coordination. Based on probabilistic information stored in memory, the decision process selects where to look next and the associated desired motor commands. Like other approaches that use uncertainty and task-related rewards to guide decision making we frame our approach within the reinforcement learning domain (Sutton and Barto 1998). As such, the agent selects the action that maximizes expected task related cumulative rewards.

The underlying framework for decision making under uncertainty, which assumes non-deterministic noisy state observations, is known as *Partially Observable Markov Decision Process* (POMDP). In our particular problem formulation we have continuous states and observations, as well as large discrete action spaces (i.e. each SES cell represents an action), which renders intractable the computation of optimal policies. Even approximate methods for solving POMDPs would take considerable time (e.g. hours or days). Moreover, if the environment changes the observer needs to recompute the full policy. Hence, state-of-the-art methods for solving POMDPs are unsuitable for large problems, which require real-time on-line decision making. Therefore, rather than framing our problem as a POMDP, we rely on simpler and less costly tools from Bayesian Optimization, for reinforcement learning. More concretely, from multi-armed bandit problems (MAB) (Robbins 1952).

6.1 Saccadic planning as a multi-armed bandit problem

In MAB problems, at each time instant the agent selects an action and collects a reward. The rewards are drawn from a posterior probability distribution whose statistics are continuously updated over time. Typically, the goal of the agent is to maximize the sum of collected rewards or, equivalently, minimize cumulative regret.

In this work the selected task was to find the closest object to the observer as fast (i.e. with minimum fixations) and precisely (i.e. with minimum uncertainty) as possible. Within the MAB framework, this is commonly referred to as the best-arm identification problem (Audibert and Bubeck 2010).

In our particular setting, each world sample point represented in memory is a bandit whose statistics are not known in advance. The agent chooses actions, i.e. a fixation point, from the set of alternatives $a \in \{1, \ldots, N_x\}$ and collects payoffs from a reward distribution $r(\mathbf{x}^a)$. Considering the task at hand, we define the reward obtained when choosing a given action a as a function of the distance to the ego-frame

$$r(\mathbf{x}^a) = -\|\mathbf{x}^a\|_2 \tag{36}$$

Since $\mathbf{x}^a$ follows a Gaussian distribution, then we consider a first order approximation for the reward distribution such that

$$r(\mathbf{x}^a) \sim \mathcal{N}\left(\mu_r(\mathbf{x}^a), \sigma_r(\mathbf{x}^a)\right) \tag{37}$$

where $\mu_r(\mathbf{x}^a)$ and $\sigma_r(\mathbf{x}^a)$ are computed as follows

$$\mu_r(\mathbf{x}^a) = E\left[-\|\mathbf{x}^a\|_2\right] = -\|\hat{\boldsymbol{\mu}}_{\mathbf{x}^a}\|_2 \tag{38}$$

$$\sigma_r(\mathbf{x}^a) = \mathrm{Var}\left[-\|\mathbf{x}^a\|_2\right] \approx \mathbf{J}^T \hat{\boldsymbol{\Sigma}}_{\mathbf{x}^a}\mathbf{J} \tag{39}$$

where $E[\cdot]$ and $\mathrm{Var}[\cdot]$ denote the expectation and variance operators, respectively, and $\mathbf{J}$ is a Jacobian matrix, defined as follows

$$\mathbf{J} = \left.\frac{\partial r(\mathbf{x})}{\partial \boldsymbol{x}}\right|_{\hat{\mu}_{\mathbf{x}^a}} = \left[\frac{x_{\hat{\mu}_{\mathbf{x}^a}}}{\|\hat{\mu}_{\mathbf{x}^a}\|_2} \; \frac{y_{\hat{\mu}_{\mathbf{x}^a}}}{\|\hat{\mu}_{\mathbf{x}^a}\|_2} \; \frac{z_{\hat{\mu}_{\mathbf{x}^a}}}{\|\hat{\mu}_{\mathbf{x}^a}\|_2}\right]^T \tag{40}$$

6.2 Acquisition functions

In the Bayesian optimization framework, acquisition functions are responsible for defining the strategy when searching for the optimum. The literature on acquisition functions used to guide stochastic optimization is vast and includes many different heuristics that deal with the exploration–exploitation dilemma. On one hand, Probability of Improvement (PI) (Kushner 1964) methods select the action that maximizes the probability of improving the current instantaneous reward. On the other hand, Expected Improvement (EI) (Mockus 1974) seeks for the action that maximizes the expected improvement magnitude. More recently, the idea of using Upper Confidence Bounds (UCB) (Lai and Robbins 1985) to deal with exploration–exploitation tradeoffs in machine learning problems has proven successful in robotics applications (Lizotte et al. 2007), exhibiting increased preference for exploration when compared to the former approaches. Since the best performing acquisition function is highly dependent on the objective at hand, the authors in Hoffman et al. (2011) propose combining single acquisition functions in mixed portfolio strategies.

In this work we compared three different action selection strategies:

1. a simple yet powerful UCB algorithm named "Sequential Design for Optimization" (Cox and John 1992) that is easy to implement and elegantly handles the trade-off between exploration (minimizing uncertainty) and exploitation (maximizing rewards) that emerges in decision making under uncertainty (Agrawal 1995). At each time instant, the observer selects the alternative with maximal upper confidence bound on the expected reward, given the past observations, according to the following expression

$$a_t = \underset{a \in \{1,\ldots,N_x\}}{\operatorname{argmax}} \; \mu_r\left(\mathbf{x}_t^a\right) + \alpha \sigma_r\left(\mathbf{x}_t^a\right) \tag{41}$$

where α is a user selected parameter that controls the width of the confidence bound and thus the exploration behaviour during task execution.
2. The probability of improvement, which at each time instant, selects the action with highest probability of leading to an improvement upon the current best ($\mathbf{x}_t^*$), as follows

$$a_t = \underset{a \in \{1,\ldots,N_x\}}{\operatorname{argmax}} \; \mathbb{P}\left(r\left(\mathbf{x}_t^a\right) > r\left(\mathbf{x}_t^*\right)\right) \tag{42}$$

3. The expected improvement, which tries to maximize the expected magnitude of the improvement upon the so far best, according to

$$a_t = \underset{a \in \{1,\ldots,N_x\}}{\operatorname{argmax}} \; \mathbb{E}\left(r\left(\mathbf{x}_t^a\right) - r\left(\mathbf{x}_t^*\right)\right) \tag{43}$$

Finally, the motor-action U_t corresponding to $\mathbf{x}_t^{a_t}$ is computed from known forward kinematics.

7 Results

In order to demonstrate the applicability of the proposed framework and compare the performance of different visual sensor topologies, we performed a set of experiments in simulation. In all of the experiments we constrained the number of RFs - and hence the computational resources - to be always fixed and equal in the Cartesian and log-polar cases (please refer to Pamplona and Bernardino 2009 for mathematical details on the log-polar distribution). We considered $N_{rf} = 200 \times 200$ images in both cases.

The remainder of this section is organized as follows. We begin by characterizing and assessing the ability of the different sensors to map the environment with low uncertainty. Then, we proceed to evaluating the performance of the complete active task-oriented stereo sensing framework, in a realistic simulated environment.

7.1 Sensor characterization

To characterize the proposed sensor model, we assessed the average uncertainty in 3D reconstruction as a function of depth, vergence angle and sensor type in the following manner: First, we generated a set of fronto-parallel planar surfaces, with varying distance $d \in [0, 1]$ from the binocular system. Depth is constant for all points lying within the same planar-surface. Then, for each planar surface we varied the vergence angle, in the interval $\theta^v \in \left[0, \frac{\pi}{2}\right]$ and computed the corresponding 3D reconstructions with associated uncertainties. Note that in this experiment we are not characterizing a full environment but just single snapshots taken by the observer. Furthermore, we assumed that an object can be approximated by a planar surface occupying the observing agent field of view. This allows for comparing both sensors, under the same conditions.

Let us consider the log-determinant of the inverse covariance matrix (also known as precision matrix) to quantify pointwise information:

$$I\left(\mathbf{\Sigma}\right) = -\log(|\mathbf{\Sigma}|) \tag{44}$$

Here we rely on the average information gathered with a single depth image to assess the quality of the sensors, which is defined as folllows

$$TI = \frac{1}{N_{rf}} \sum_{i=1}^{N_{rf}} I\left(\hat{\mathbf{\Sigma}}^i\right) \tag{45}$$

As depicted in Fig. 5, the foveal outperforms the Cartesian sensor, in terms of gathered information which is maximal if the fixation point coincides with the planar surface. Furthermore, the Cartesian sensor information reliability does not depend on the vergence angle and decays monotonically with increasing depth, while the foveal sensor performance critically depends on the careful selection of the vergence angle. These results are directly in line and support previous findings (Weiman 1995) that suggest that foveal distributions facilitate stereo vision in convergent systems. In foveated systems gaze acts like a focus of attention, which when directed to the point of interest, improves dramatically the depth resolution around the fixation point. Instead, optical vergence

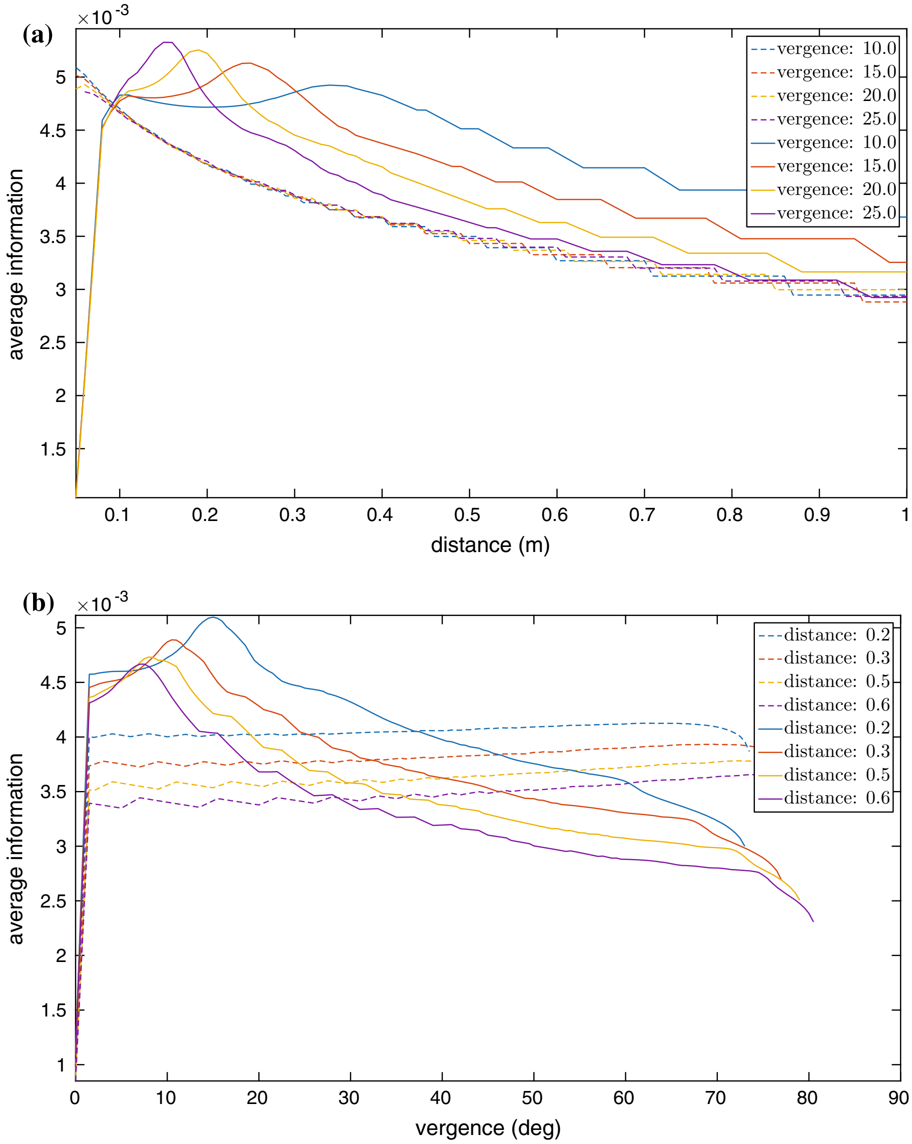

Fig. 5 Numerical characterization of the sensor model for the Cartesian (*dashed lines*) and the log-polar (*solid lines*) sensors, as a function of distance and vergence. **a** Varying distance for diferent vergence angles curves and **b** varying vergence for different planar distance curves

movements in Cartesian systems provide no gains in 3D resolution, resulting only in unnecessary energetic costs.

7.2 Active vision

With the view of investigating how our methodology performs in simultaneous target searching and mapping, we performed a set of experiments in the Gazebo simulator with the Vizzy robot (Moreno et al. 2015) head (see Fig. 6). For the sake of the experiments simplicity, the robot was fixed to the ground floor and hence the motion was restricted to oculocephalic movements. However, note that our methodology is also applicable to scenarios in which the robot platform can move. This would imply updating the 3D point estimates stored in memory taking into account the uncertainty in robot base movements (odometry), and implementing a z-buffer technique to determine which point to store in each cell, due to possible occlusions occurring after translations.

We created a static scenario with multiple objects (coke cans) displaced at arbitrary depths, over a highly textured background, in order to facilitate stereo reconstruction, which is highly dependent on the environment texture richness. The Gazebo simulator generates pinhole camera images, with uniform resolution. Hence, for the log-polar sensor, we generated foveated images from uniform resolution images by first applying the log-polar transformation and then converting back to Cartesian domain via the inverse transformation. This operation has the effect of blurring the image in the periphery while maintaining high resolution in the center. Finally, disparity maps were computed using a state-of-the-art dense stereo matching algorithm named Semi-Global Block Matching (SGBM) (Hirschmuller 2008).

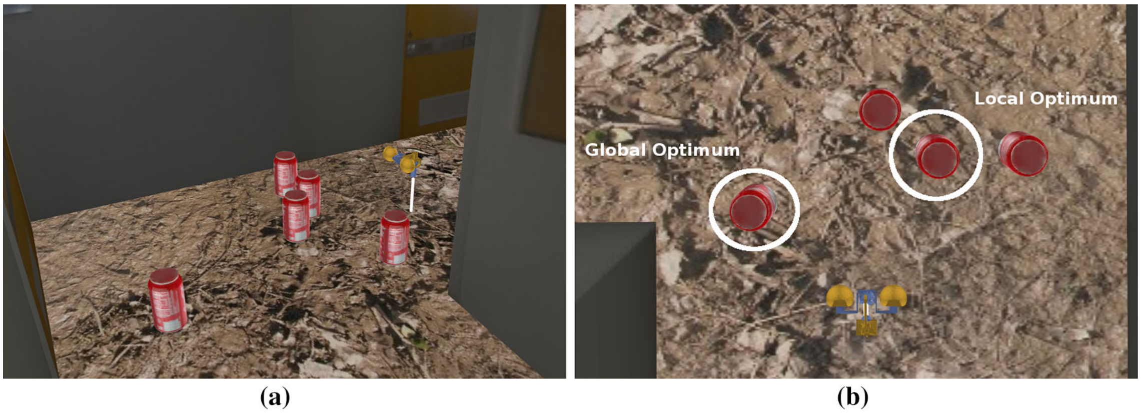

Fig. 6 The simulation scenario created for evaluating the proposed active vision framework. The task to perform was to find the nearest object from the robot ego frame. The evaluation scenario contained a non-trivial global optimum which could only be attended if enough exploration was promoted. **a** The simulation scenario created for evaluating the proposed active vision framework and **b** the global optimum was placed at a non-trivial location which could only be attended with either sufficient exploration or a wide field of view

As previously pointed out, the task at hand was to find the nearest world point to the observer. Points on the ground floor are easily excluded by thresholding the z_w coordinate. In all experiments we fixed the number of memory sample points to $N_x = 20,000$. In each experiment we let the observer perform $T = 50$ saccadic movements, with initial ($t = 1$) pan, tilt and vergence angles equal to zero. Each experiment was repeated 20 times in order to average out variability in different real-time simulations. Non-repeatability was influenced by multiple factors including separate threads for Gazebo's physics and sensor generation, as well as stochastic delays involved in higher level inter-process communication. Furthermore, in order to deal with motion blur and visuo-proprioceptive delays that arise during saccadic eye movements, we used the visual suppression mechanism proposed in Avelino et al. (2016), which temporarily blinds the observer during saccades.

7.2.1 Evaluation metrics

In order to quantitatively assess the performance of our methodologies we considered the following evaluation metrics:

- the gap reduction metric (Huang et al. 2006) which is a quality measure that evaluates how effectively the algorithm is at finding the global maximum:

$$g_t = \frac{\mu_r(\mathbf{x}^+) - \mu_r(\hat{\mathbf{x}}_1^{a_1})}{\mu_r(\mathbf{x}^*) - \mu_r(\hat{\mathbf{x}}_1^{a_1})} \tag{46}$$

where $\mu_r(\mathbf{x}^*)$ is the true global maximum

$$\mu_r(\mathbf{x}^*) = \max_i \mu_r(\mathbf{x}^i) \tag{47}$$

and $\mu_r(\mathbf{x}^+)$ is the best obtained reward up to time t

$$\mu_r(\mathbf{x}^+) = \max_t \mu_r(\hat{\mathbf{x}}_t^{a_t}) \tag{48}$$

The gap is defined between 0, meaning no improvement over the initial fixation, and 1 for the optimal improvement. In order to measure the speed for task completion and thus performance efficiency we also assess the average gap reduction per saccade which implicitly represents the average progress towards the optimum per saccade:

$$G/S = \frac{1}{T}\sum_{t=1}^{T} g_t \tag{49}$$

- the cumulative regret which is a standard metric, here suitable to evaluate the convergence behaviour during the search for the optimum:

$$R_t = \mu_r(\mathbf{x}^*) - \frac{1}{t}\sum_{k=1}^{t} \mu_r(\hat{\mathbf{x}}_k^{a_k}) \tag{50}$$

Notice that here we are not interested in minimizing the total regret, i.e. the incurred losses during exploration, but instead on finding the global optimum. When normalized by the number of saccades it represents the temporal cumulative regret gain per saccade:

$$R/S = \frac{1}{T}\sum_{t=1}^{T} R_t \tag{51}$$

- the average global gathered information which is a quality performance measure of the global knowledge gathered about the world up to time t (exploratory behaviour):

Table 1 Memory biasing parameters

Bias	$\boldsymbol{\mu}_p$			$\boldsymbol{\Sigma}_p$		
	x	y	z	xx	yy	zz
Unbiased	0	0	0	0.5	0.5	0.5
Top	0	0	1	0.5	0.5	0.5
Down	0	0	−1	0.5	0.5	0.5
Target	0.61	0.43	−0.67	0.05	0.05	0.05

$$GI_t = \frac{1}{N_x} \sum_{i=1}^{N_x} I\left(\hat{\boldsymbol{\Sigma}}_t^i\right) \tag{52}$$

When normalized by the number of saccades it represents the temporal average global information gain per saccade:

$$GI/S = \frac{1}{T} \sum_{t=1}^{T} GI_t \tag{53}$$

- the nearest object gathered information, which is a target reconstruction quality measure that benefits high precision (i.e. low uncertainty) in target reconstruction:

$$LI_t = I\left(\hat{\boldsymbol{\Sigma}}_t^i\right) \quad \forall_{i:\|\hat{\mu}_{\mathbf{x}_t^*}-\hat{\mu}_{\mathbf{x}_t^i}\|<R_{NN}} \tag{54}$$

where $\hat{\boldsymbol{\Sigma}}_t^*$ is the true known global maximum estimated covariance at time t and R_{NN} is a user-selected nearest-neighbor radius. We considered $R_{NN} = 0.1m$ in all the experiments described below.
When normalized by the number of saccades it represents the temporal average local information gain per saccade:

$$LI/S = \frac{1}{T} \sum_{t=1}^{T} LI_t \tag{55}$$

7.2.2 Foveal versus Cartesian

Our first aim was to compare the behaviour of the foveal against the Cartesian sensor during task execution, for different upper confidence bound parameter values $\alpha \in \{0, 0.01, 1, 100, \infty\}$ and different sensing field of views fov $\in \{90, 135\}$. The sensor field of views were selected such that in one of the cases (fov $= 90$) the global optimum was not in the field of view of the observer at $t = 1$. The SES cells were generated from a unbiased, zero mean Gaussian distribution at initialization (see Table 1).

A global analysis of the results depicted in Fig. 8 shows that the foveal sensor outperforms the Cartesian both in terms of the quality of the gathered information, as well as the task execution speed and effectiveness, as demonstrated by the gap reduction plot. We hypothesize that the best performance of the foveal sensor is due to the fact that the uncertainty in the periphery implicitly promotes more peripheral (lateral) exploration whereas the Cartesian promotes longitudinal (depth) search. This statement is clearly supported by the cumulative regret plots which exhibit lower losses for the Cartesian sensor, and thus a greedier behaviour. Moreover, for the foveal sensor case, a larger FOV allows the agent to attend the target more quickly at the cost of reduced information gain. A wider FOV, despite having less peripheral resolution, is advantageous in the speed of execution during visual search tasks (Fig. 7).

In Fig. 8a we assess the performance of our method for the different acquisition functions referred in Sect. 6. On one hand, in the UCB case, a larger confidence bound parameter α increases exploration and, on average, improves performance in the particular task of finding the nearest object. However, too much exploration incurs in large cumulative regrets, and thus in high energy costs due to large oculo-cephalic movements when attending objects further from the observer. Nevertheless, purely exploratory behaviours ($\alpha = \infty$) lead to better results in the average reconstruction quality as shown by the information metrics, since on average more memory sample locations are fixated. On the other hand, the tested improvement-based policies (PI and EI) seek to improve on the current best and have the advantage of being parameter free. For our particular setting, and similarly to UCB with $\alpha = 0$, PI tends to be excessively greedy and get trapped in local minima. On the contrary, EI deals well with the exploration–exploitation trade-off, as demonstrated by the average gap reduction and cumulative regret per saccade metrics due to the fact that it implicitly accounts for the improvement magnitude of each saccadic action, which allows for choosing distant, with high variance, fixation points.

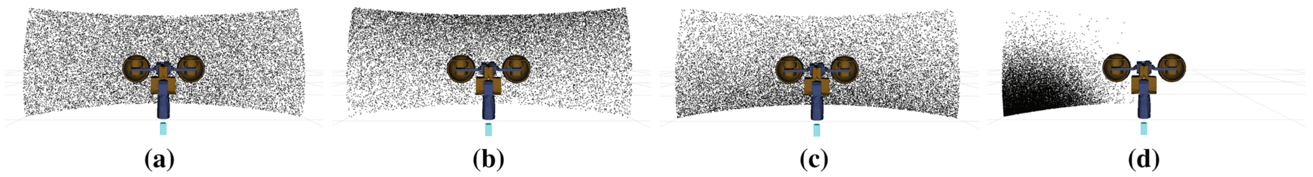

Fig. 7 SES sample point distribution according to different topological memory biases and kinematic constraints. **a** Uniform, **b** top bias, **c** bottom bias and **d** target bias

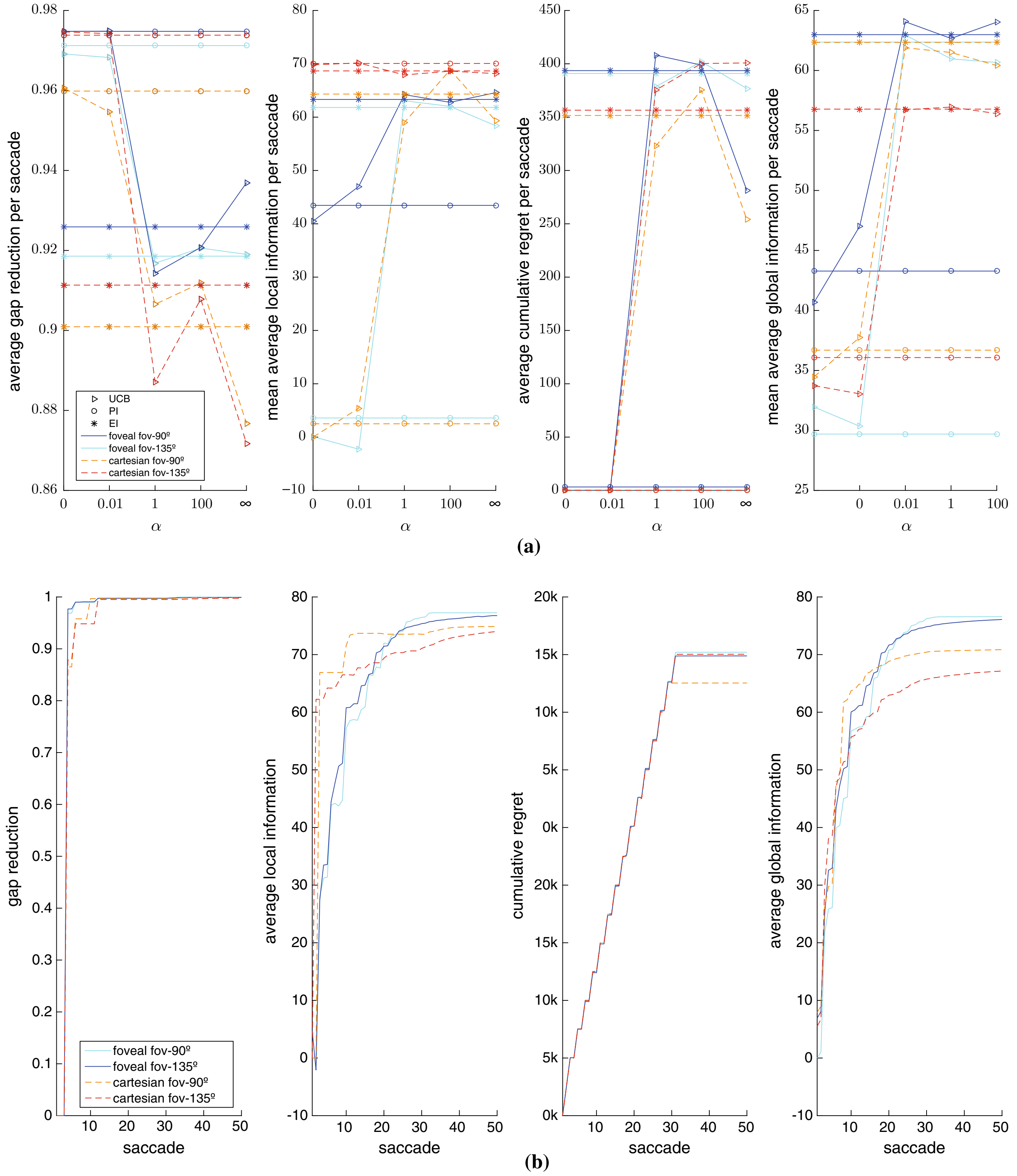

Fig. 8 Performance results for the assessed sensor topologies, field of views and upper confidence bound parameter. **a** Average per saccade performance plots and **b** time evolution plots

An in-depth analysis of the temporal evolution metrics (Fig. 8b) for a fixed $\alpha = 100$, allows us to assess convergence times for a fairly exploratory behaviour. The temporal evolution of the gap reduction metric shows that, in all cases, no more than 20 saccades are necessary to perform the task of finding the nearest object for both sensor types. Howbeit,

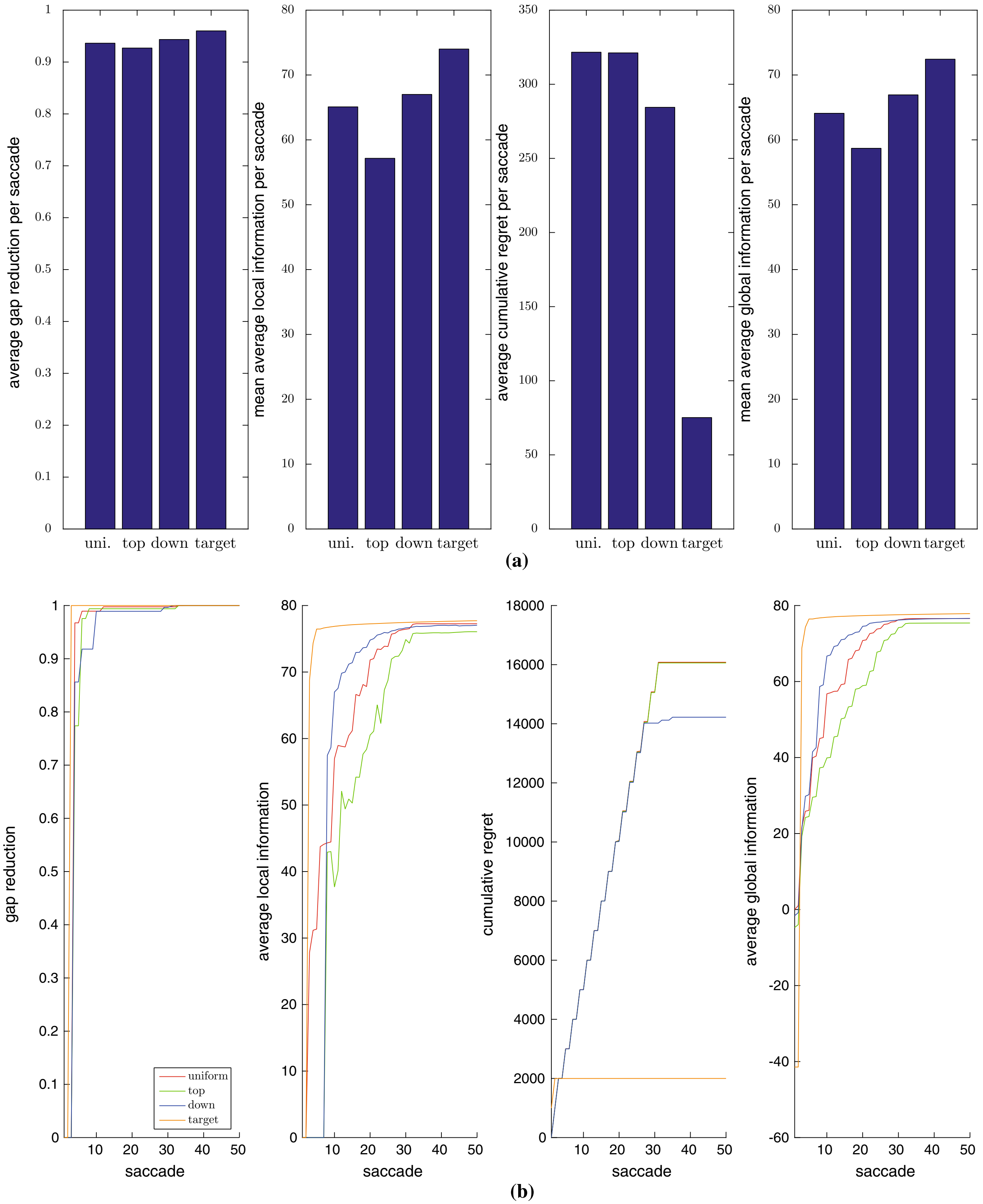

Fig. 9 Performance results for the assessed memory biases. **a** Average per saccade performance plots $\alpha = 100$ and **b** time evolution plots

as indicated by the accumulated regret temporal evolution, convergence is only achieved after no less then 30 saccades. We further note that, after convergence, the cumulative regret is on average higher for the foveal case, as a consequence of having a more exploratory nature. The gathered information exhibits an asymptotically convergence behaviour and has a faster transient time for the Cartesian sensor, again, supporting the idea that the Cartesian sensor is more greedy, myopic, and thus more prone to get trapped in local minima.

7.2.3 Memory biases

Here our goal was to investigate the effect of different spatial memory topological biases imposed from a priori knowledge regarding the environment structure and the task at hand. At the present, experiments were performed with the foveal sensor with a fov $= 135$, and for the UCB with $\alpha = 100$. We intended to demonstrate that a careful displacement of the memory patches considering prior knowledge about the surrounding environment and the task at hand should incur in large performance gains. Therefore, we considered four different prior belief distributions with parameters defined in Table 1 and resulting SES topologies depicted in Fig. 7:

- a “neutral” (unbiased) distribution reflecting the absence of a priori knowledge about the target location.
- a “bad” (top) prior belief distribution based on the wrong assumption that the object is above the observer.
- a “good” (down) prior belief distribution that assumes that the object is on the ground
- a “very good” (target) prior belief considering the true location of the target object.

In the Fig. 9 we can observe that the “target” case had the best performance and the “top” the worst performance according to all metrics. In fact, as demonstrated by the gap reduction and the accumulated regret time evolution plots, the method was successful in finding the global optimum and converged with only 2 saccades. All the other cases were still able to find the optimum with less than 10 saccades and converge to the optimum within the first 20 saccades.

As expected, the gathered average local information metric indicates that increasing the memory sample density around the object of interest improves the target’s gathered information. These experiments demonstrate that translating task-related priors in clever memory allocation to regions of higher reward yields faster task execution times and faster convergence rates. This results in an increase in the time spent on reducing the uncertainty on the target and therefore in improved reconstruction quality. On the one hand, promoting higher resolution in spatial memory to the most important surrounding regions according to the task, allows for more accurate target reconstruction. On the other hand, less fixations are needed to find the target, since less memory cells, and thus possible fixations, will reside outside of the target vicinity.

8 Conclusion

In this work we investigated the impact of uncertainty due to quantization phenomena in the retina and on how to take advantage of it to guide gaze shifts for two distinct retinal topologies: Cartesian and log-polar. With our approaches different gaze patterns emerge depending on the sensor topology and field of view and on exploration–exploitation confidence bounds parameters. The obtained results demonstrate that a wider field of view, despite less peripheral resolution is advantageous in visual search tasks execution speed. Furthermore, we showed that a task-biased SES allows for simultaneously coping with limited memory resources (i.e. limited number of memory cells) while improving performance, both in terms of target reconstruction quality and task execution speed.

The proposed framework can be further enhanced with other ideas from the attentional stereopsis literature, namely with (Agarwal and Blake 2010) which improves reconstruction quality and efficiency by restricting stereo matching to biologically plausible volumes of interest. Moreover, one could improve the proposed SES run-time performance by relying on a nearest neighbour data alignment scheme. A kd-tree could be built during initialization time and be used for storing the ego-sphere cells ($\mathcal{P}$). Then, for each observed 3D point, searching for the closest cell on the sphere would be performed in $O(\log N_x)$, instead of $O(N_x)$. Future work will include developing adaptive re-sampling techniques for online memory biasing, capable of coping with dynamic tasks and environments.

Acknowledgements This work has been partially supported by the Portuguese Foundation for Science and Technology (FCT) Project [UID/EEA/50009/2013]. Rui Figueiredo is funded by FCT Ph.D. Grant PD/BD/105779/2014. Helder Araújo would like to thank FCT (Portuguese Foundation for Science and Technology) grant UID-EEA-0048-2013.

References

Agarwal, A., & Blake, A. (2010). Dense stereo matching over the panum band. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(3), 416–430.

Agrawal, R. (1995). Sample mean based index policies with o (log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, *27*, 1054–1078.

Ahmad, S., & Yu, A. J. (2013). Active sensing as bayes-optimal sequential decision making, *CoRR*, vol. abs/1305.6650. http://arxiv.org/abs/1305.6650.

Audibert, J. -Y., & Bubeck, S. (2010). Best arm identification in multi-armed bandits. In *COLT-23th conference on learning theory-2010* (pp. 13-p).

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, *47*(2–3), 235–256.

Avelino, J. A., Figueiredo, R., Moreno, P., & Bernardino, A. (2016). On the perceptual advantages of visual suppression mechanisms for dynamic robot systems. In *International conference on biologically inspired cognitive architectures (BICA)*.

Begum, M., & Karray, F. (2011). Visual attention for robotic cognition: A survey. *IEEE Transactions on Autonomous Mental Development*, *3*(1), 92–105.

Bernardino, A., & Santos-Victor, J. (2002). A binocular stereo algorithm for log-polar foveated systems. In H. Blthoff, C. Wallraven, S. -W. Lee , & T. Poggio (Eds.), *Biologically motivated computer vision, ser. Lecture notes in computer science*, (Vol. 2525, pp. 127–136). Berlin: Springer.

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 185–207.

Butko, N. J., & Movellan, J. R. (2010). Infomax control of eye movements. *IEEE Transactions on Autonomous Mental Development*, *2*(2), 91–107.

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision research*, *51*(13), 1484–1525. (vision Research 50th Anniversary Issue: Part 2). http://www.sciencedirect.com/science/article/pii/S0042698911001544.

Colombo, C., Rucci, M., & Dario, P. (1996). Integrating selective attention and space-variant sensing in machine vision. In J. L. C. Sanz (Ed.), *Image technology: Advances in image processing, multimedia and machine vision* (pp. 109–127). Springer Berlin Heidelberg.

Cox, D. D., John, S. (1992). Sdo: A statistical method for global optimization. In *IEEE international conference on systems, man and cybernetics* (pp. 1241–1246). IEEE.

Crawford, L. E., Landy, D., & Presson, A. N. (2014). Bias in spatial memory: Prototypes or relational categories. In *Poster presented at the 36th annual conference of the cognitive science Society, Quebec*.

Edelman, S. (1995). Receptive fields for vision: From hyperacuity to object recognition. http://cogprints.org/570/.

Ferreira, J., Bessière, P., Mekhnacha, K., Lobo, J., Dias, J., & Laugier, C. (2008). Bayesian models for multimodal perception of 3D structure and motion. In *International conference on cognitive systems (CogSys 2008)*, Karlsruhe, Germany. https://hal.archives-ouvertes.fr/hal-00338800.

Fleming, K. A., Peters, R. A., & Bodenheimer, R. E. (2006). Image mapping and visual attention on a sensory ego-sphere. In *2006 IEEE/RSJ international conference on intelligent robots and systems, IROS 2006, Beijing, China* (pp. 241–246). October 9-15, 2006. doi:10.1109/IROS.2006.281688.

Friston, K., Adams, R., & Montague, R. (2012). What is value accumulated reward or evidence? *Frontiers in Neurorobotics*,. doi:10.3389/fnbot.2012.00011.

Hirose, M., Furuhashi, H., Miyasaka, T., & Araki, K. (2002). Reconstruction of range data by means of geodesic dome type data structure. *The Journal of the Institute of Image Electronics Engineers of Japan*, *31*(3), 388–395.

Hirschmuller, H. (2008). Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, *30*(2), 328–341.

Hoffman, M. D., Brochu, E., & de Freitas, N. (2011). Portfolio allocation for bayesian optimization. Citeseer.

Hornung, A., Wurm, K. M., Bennewitz, M., Stachniss, C., & Burgard, W. (2013). OctoMap: An efficient probabilistic 3D mapping framework based on octrees. *Autonomous Robots*. http://octomap.github.com.

Huang, D., Allen, T. T., Notz, W. I., & Zeng, N. (2006). Global optimization of stochastic black-box systems via sequential kriging meta-models. *Journal of Global Optimization*, *34*(3), 441–466.

Itti, L., & Baldi, P. F. (2006). Bayesian surprise attracts human attention. In *Advances in neural information processing systems (NIPS*2005)* (Vol. 19, pp. 547–554). Cambridge, MA: MIT Press. su;mod;bu;td;ey.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*, 1254–1259.

Julier, S., & Uhlmann, J. (2004). Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, *92*(3), 401–422.

Koch, C., & Ullman, S. (1987). Shifts in selective visual attention: Towards the underlying neural circuitry. In L. M. Vaina (Ed.), *Matters of intelligence: Conceptual structures in cognitive neuroscience* (pp. 115–141). Dordrecht: Springer Netherlands.

Kriegman, D. J., Triendl, E., & Binford, T. O. (1989). Stereo vision and navigation in buildings for mobile robots. *IEEE Transactions on Robotics and Automation*, *5*(6), 792–803.

Kushner, H. J. (1964). A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *Journal of Fluids Engineering*, *86*(1), 97–106.

Lai, T. L., & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, *6*(1), 4–22.

Lizotte, D., Wang, T., Bowling, M., & Schuurmans, D. (2007). Automatic gait optimization with gaussian process regression. In *Proceedings of the IJCAI* (pp. 944–949).

Mockus, J. (1974). On bayesian methods for seeking the extremum. In *Proceedings of the IFIP technical conference* (pp. 400–404). London, UK: Springer. http://dl.acm.org/citation.cfm?id=646296.687872.

Moreno, P., Nunes, R., Figueiredo, R., Ferreira, R., Bernardino, A., Santos-Victor, J., Beira, R., Vargas, L., Aragão, D., & Aragão, M. (2015). Vizzy: A humanoid on wheels for assistive robotics. In *Robot 2015: Second Iberian robotics conference* (pp. 17–28). Springer International Publishing 2016.

Muller, M. E. (1959). A note on a method for generating points uniformly on n-dimensional spheres. *Communications of the ACM*, *2*(4), 19–20. doi:10.1145/377939.377946.

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*(7031), 387–391.

Pamplona, D., & Bernardino, A. (2009). Smooth foveal vision with Gaussian receptive fields. In *9th IEEE-RAS international conference on humanoid robots, humanoids 2009, Paris, France* (pp. 223–229). December 7–10, 2009. http://dx.doi.org/10.1109/ICHR.2009.5379575.

Perrollaz, M., Spalanzani, A., & Aubert, D. (2010). Probabilistic representation of the uncertainty of stereo-vision and application to obstacle detection. In *Intelligent vehicles symposium (IV), 2010 IEEE* (pp. 313–318). June 2010.

Peters, R. A., Hambuchen, K. A., & Bodenheimer, R. E. (2009). The sensory ego-sphere: A mediating interface between sensors and cognition. *Autonomous Robots*, *26*(1), 1–19. doi:10.1007/s10514-008-9098-3.

Posner, M. (2012). Cognitive neuroscience of attention. Guilford Press. http://books.google.pt/books?id=8yjEjoS7EQsC.

Robbins, H., et al. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, *58*(5), 527–535.

Ruesch, J., Lopes, M., Bernardino, A., Hornstein, J., Santos-Victor, J., & Pfeifer, R. (2008). Multimodal saliency-based bottom-up attention a framework for the humanoid robot ICUB. In *IEEE international conference on robotics and automation, 2008. ICRA 2008* (pp. 962–967). May 2008.

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning* (1st ed.). Cambridge, MA: MIT Press.

Tippetts, B., Lee, D. J., Lillywhite, K., & Archibald, J. (2016). Review of stereo vision algorithms and their suitability for resource-limited systems. *Journal of Real-Time Image Processing*, *11*(1), 5–25.

Vijayakumar, S., Conradt, J., Shibata, T., & Schaal, S. (2001). Overt visual attention for a humanoid robot. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems, 2001* (Vol. 4, pp. 2332–2337). IEEE.

von Helmholtz, H. & König, A. (1896). *Handbuch der physiologischen Optik* (Vol. 1). L. Voss. https://books.google.pt/books?id=Lb4KAAAAIAAJ.

Wang, J., & Liu, Y. (2007). A closed-form solution of reconstruction from nonparallel stereo geometry used in image guided system for surgery. In N. Sebe, Y. Liu, Y. Zhuang, & T. Huang (Eds) *Multimedia content analysis and mining*, ser. Lecture notes in computer science (Vol. 4577, pp. 371–380). Berlin Heidelberg: Springer.

Weiman, C. F. R. (1995). Binocular stereo via log-polar retinas. In *SPIE*, Ed.

Rui Pimentel de Figueiredo received a M.Sc. degree in Electrical and Computer Engineering from Instituto Superior Técnico(IST), Lisbon, Portugal, in 2012. He has been a Research Assistant member of the Computer Vision Laboratory (VisLab), Institute for Systems and Robotics (ISR), Lisbon. His work as has been related with 3D geometry processing, computer vision and robotics, with a strong emphasis on the subject of object recognition. He has been responsible for software development, implementation and maintenance within two EU projects (First-MM and HANDLE) which were mainly directed towards robot grasping and in-hand manipulation applications. He is a Ph.D. student since 2014. His main research goal is to understand and model the impact of non-uniform resolution images in the behavior of biological and artificial systems during visual-search tasks.

Alexandre Bernardino is an Associate Professor at the Department of Electrical and Computer Engineering of IST-Lisboa and Senior Researcher at the Computer and Robot Vision Laboratory of the Institute for Systems and Robotics of IST-Lisboa. He has participated in several national and international research projects as principal investigator and technical manager. He published more than one hundred research papers on top journals and peer-reviewed conferences in the field of robotics, vision and cognitive systems. He is associate editor of the journal Frontiers in Robotics and AI and of major robotics conferences. He is the chair or the IEEE Portugal RAS Chapter. His main research interests focus on the application of computer vision, machine learning, cognitive science and control theory to advanced robotics and automation systems.

José Santos-Victor is a Full Professor of Electrical and Computer Engineering at IST, Lisbon. He is currently President of the Institute for Systems and Robotics (ISR) where he founded the the Computer and Robot Vision Lab (VisLab). His main research interests are in the area of vision-based control and navigation, and cognitive robots and systems. He has been the PI for IST in many EU Projects dealing with robotics, vision, neuroscience and developmental psychology. He coordinates an international Ph.D. Program in Robotics, Brain and Cognition. He supervised 15 Ph.D. students and published over 40 journal papers in the area of computer vision and robotics. Between 2006 and 2014, he served as IST Vice President for International Affairs, developing academic collaboration activities in different areas of the globe, namely with Brazil, Africa and China, besides Europe.

Helder Araújo is a Professor in the Department of Electrical and Computer Engineering of the University of Coimbra, Portugal and also an Invited Professor at Blaise Pascal University, Clermont-Ferrand, France. He is a researcher at the Institute for Systems and Robotics Coimbra. His research interests include computer and robot vision, robot navigation and localization, and sensor modeling for robot navigation and localization.