Motor-primed Visual Attention for Humanoid Robots

Luka Lukic, Aude Billard, and José Santos-Victor

***Abstract*—We present a novel, biologically-inspired, approach to an efficient allocation of visual resources for humanoid robots in a form of a motor-primed visual attentional landscape. The attentional landscape is a more general, dynamic and a more complex concept of an arrangement of spatial attention than the popular "attentional spotlight" or "zoom-lens" models of attention. Motor-priming of attention is a mechanism for prioritizing visual processing to motor-relevant parts of the visual field, in contrast to other, motor-irrelevant, parts. In particular, we present two techniques for constructing a visual "attentional landscape". The first, more general, technique, is to devote visual attention to the reachable space of a robot (peripersonal space-primed attention). The second, more specialized, technique is to allocate visual attention with respect to motor plans of the robot (motor plans-primed attention). Hence, in our model, visual attention is not exclusively defined in terms of visual saliency in color, texture or intensity cues, it is rather modulated by motor information. This computational model is inspired by recent findings in visual neuroscience and psychology. In addition to two approaches to constructing the attentional landscape, we present two methods for using the attentional landscape for driving visual processing. We show that motor-priming of visual attention can be used to very efficiently distribute limited computational resources devoted to the visual processing. The proposed model is validated in a series of experiments conducted with the iCub robot, both using the simulator and the real robot.**

***Index Terms*—Cognitive robotics, computer vision, humanoid robots, machine learning.**

I. Introduction

VISION is one of the most important functional modules in both artificial and biological systems, and yet one of the most computationally demanding ones. In a humanoid robot, the computational demands for processing stereo images represent very often a bottleneck for real-time manipulation, where replanning and computation of visuomotor actions are time-locked within a time range of only a few milliseconds. Most of the approaches in robot vision are based on the standard, "off the shelf", image processing techniques, ignoring most, if not all, the information regarding the current motor state and planned motor actions. This implies that the visual system and the arm-hand system are usually considered as two largely independent modules that communicate only in the direction from vision to manipulation, which implies that during visual processing the valuable information from the manipulation system is mostly ignored. This decoupling of visual processing from the motor information manifests itself in an inefficient, hence slow, visual processing. In this work we show that modulation of visual processing, which emerges from the motor system, can drastically improve visual performances, in particular, the speed of visual computation, one of the most critical aspects of the system. Fig. 1 illustrates the main principles of this work.

If we put this in a real-world context, let us imagine a robot bartender, equipped with an active stereo camera system that

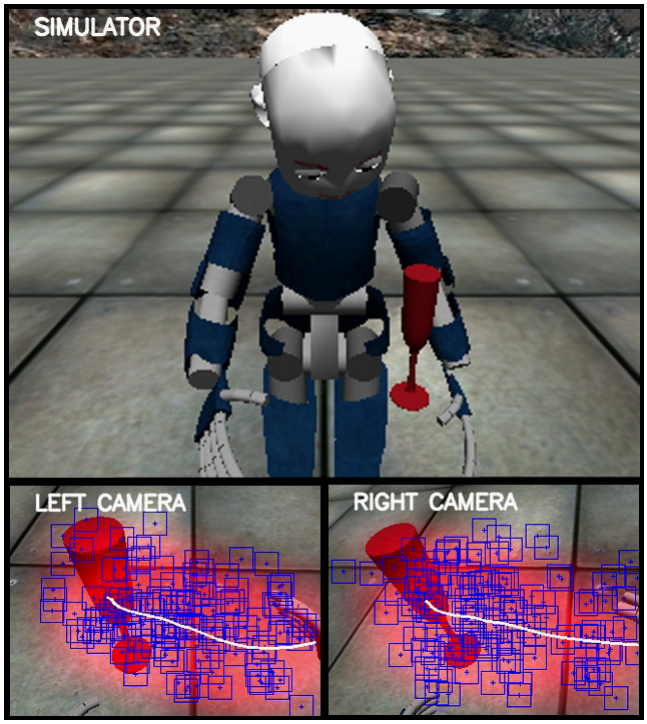


Figure 1. The figure displays the main idea of the proposed approach: nonuniform image processing driven by a motor-primed visual attentional landscape. Visual space is prioritized depending on its motor relevance, i.e., visual attention is biased toward motor-relevant parts of the workspace projected to the stereo images. The white line represents a forward-planned (mentally-simulated) movement toward the object to be grasped (red glass). The reddish blend superimposed on the snapshots of the left and right cameras is a visualization of the intensity of the visual attentional landscape. The attentional landscape has a higher intensity closer to motor relevant parts of the visual field. The images are processed in a manner that the spatial distribution of their attentional landscapes is taken into account (motor-relevance is prioritized). The anchors of the scanning windows (blue squares) are sampled with respect to their relevance, i.e. more dense visual scanning is done where the attentional landscape has higher values, and less dense scanning where it has low values. Ignoring irrelevant parts of the images affords significant computational savings, whereas the processing of motor-relevant parts of the visual scene supports visually-guided reaching and grasping.

This work was partially supported by the FCT (PEst-OE/EEI/LA0009/2013), RBCog – Robotics, Brain and Cognition doctoral program, and the EU project POETICON++ (FP7-ICT-288382). The work of L. Lukic was supported by a PhD Student Scholarship from the FCT, SFRH/BD/51072/2010 under the IST-EPFL Joint Doctoral Initiative.

Luka Lukic and José Santos-Victor are with VISLAB/ISR, Instituto Superior Técnico, Lisbon, Portugal (e-mail: luka.lukic@epfl.ch, jasv@isr.ist.utl.pt).

Aude Billard is with LASA, EPFL, Lausanne, Switzerland (e-mail: aude.billard@epfl.ch)

has the task to grasp a glass, fill it with a beverage of choice, and serve it to a guest. In a visually-aided manipulation, based on the standard vision processing approach, during reaching and grasping for the target object, in every cycle of the control loop, vision scans every part of both stereo images searching for the target object and potential obstacles, in order to update the robot's knowledge about their state (position, orientation and other properties of interest that might change during a task). Assume that the motion of the arm has been initiated and is directed toward a specific object, say a wine glass (the obstacles will by definition be all objects that obstruct an intended movement). Here, a question arises: why would one want to scan the peripheral parts of the stereo images for obstacles, since these correspond to regions in the workspace ten meters or so from the wine glass that is at around 30 cm from the hand? Clearly, space scanned should be restricted to a region of space that is motor-relevant.

Contrary to robots, humans and non-human primates have the ability to rapidly and graciously perform complicated tasks with a limited amount of computational resources. One of the reasons for their superior performances in visuomotor tasks is an efficient distribution of the visual resources to select only relevant information for reaching and grasping among the plethora of visual information. Humans are able to efficiently and routinely manage this challenging task of selective information processing, in a seemingly effortless manner, by means of highly customized attentional mechanisms. In dynamically changing environmental conditions, the time pressure for rapid computation cannot afford the computation of the full-blown visual model of the world [1], [2]. For this reason, humans and non-human primates use the attentional mechanisms to select important visual information, and cheaply compute only a relevant subset of them on the fly. In visual attention, two mechanisms are recognized: *covert attention* and *overt attention* [3]. Covert visual attention corresponds to an allocation of mental resources for processing extrafoveal visual stimuli. Overt visual attention consists in active visual exploration involving saccadic eye movements (Fig. 2). These two mechanisms are instantiations of the same underlying mechanism of visual attention, hence intermingled both functionally and structurally, working in synchronization and complementing each other. Covert attention selects interesting regions in the visual field, which are subsequently attended with overt gaze movements for high-acuity foveated extraction of information [4], [5], [6]. Furthermore, visual attention (covert and overt) is tightly coupled with the motor system. Numerous findings from visual neuroscience and psychology provide evidence that visual attention is bound and actively tailored with respect to spatio-temporal requirements of manipulation tasks [7], [8], [9], [10], [11]. Fig. 2 illustrates how attention is drawn toward manipulation-relevant regions of the visual field, even in a common, well-rehearsed natural task such as tea serving.

In this paper, we hypothesize that such a biologically-inspired, explicit, active adaptation of attention with respect to motor plans can endow robot vision with a mechanism for the efficient allocation of limited visual resources. This approach contributes to the state of the art in visual-based reaching and grasping, tackling visual attention from a new, alternative perspective where visual attentional relevance is not defined in terms of low-level visual features such as color, texture or intensity of the visual stimuli, but rather in terms of manipulation-relevant parts of the visual field as visually relevant regions. In our model, the attentional mechanism becomes a fundamental building element of the motor planning system and vice versa. At each cycle of the control loop, the visual and motor systems modulate each other sending each other control signals. The proposed approach is evaluated in robotic experiments using the iCub humanoid robot [16].

The work reported in this paper was published in a preliminary form in [17]. The present paper extends the previous work in three ways: (a) we develop a novel model of peripersonal space-primed attention, (b) we improve the previously proposed sampling scheme for attentional driving of visual processing, and (c) we verify the presented approach in more robot experiments.

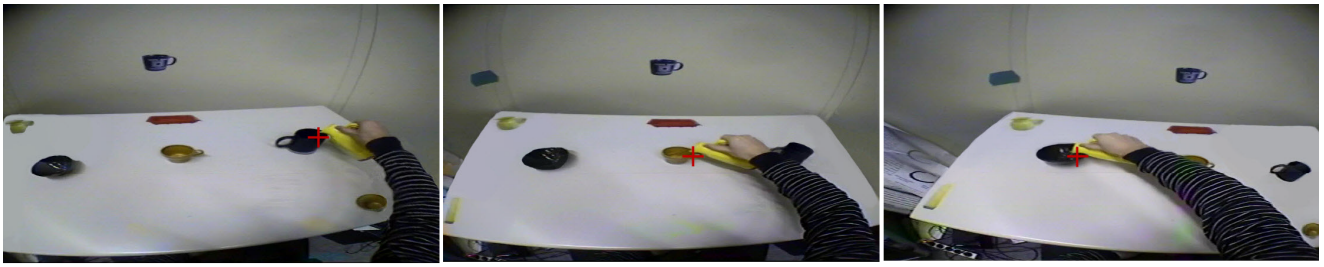
The rest of the paper is organized as follows. In Section II, we briefly review related work on computational modeling of visual attention, its use in robotics, and the biological evidence onto which we ground our approach to tackle the existing problems. In Section III, we present our two approaches for obtaining the attentional landscape. Section IV describes how image processing is performed once the attentional landscape is computed. Section V reports on validations of the approach in experiments with the iCub robot. Section VI is devoted to discussion.

II. Related work

A. Computational modeling of attention

Most of the modern work on computational modeling of attention draws inspiration from the feature integration theory of attention from psychology [18]. The feature integration theory argues that low-level, pre-attentive features attract visual attention in a bottom-up, task-independent manner. The intuition behind this approach is that a non-uniform spatial distribution of features is somehow correlated with their informative significance. The influence of the low-level features on capturing attention is motivated by the functions of the neural circuitry in the early primate vision and experimental findings in scene observation tasks [19], [20], [10]. By far, the most influential computational implementation grounded in this theory is the concept of the saliency map [21]. In the aforementioned model, low-level features such as color, orientation, brightness and motion are extracted in parallel from the visual input. The visual input is represented as a digitized 2D image. Low-level features from the visual stimuli compete across local neighborhoods and multiple spatial scales building spatial banks of features that correspond to center-surround contrast computed across different scales. The feature banks are normalized and aggregated by a weighted sum to create a master saliency map. The focus of attention is driven by the interplay between a winner-take-all mechanism (WTA) and an inhibition of return mechanism (IOR) that operates on the final saliency map.

This pure bottom-up, approach driven by the early perceptual pop-out features has been subsequently extended to



(a) (b) (c)

Figure 2. Experimental setup with a natural task performed by the authors to illustrate the interaction between overt visual attention (gaze movements) and the arm motor system. The subject is required to pour the tea into two cups and one bowl that are placed close to the horizontal midline of the table. 4 pictures of various objects are placed close to the border of the table and 2 pictures are placed on the wall facing the subject. These pictures play the role of visually salient distractors because they share the same visual features with the objects, but remain completely irrelevant for manipulation through the entire task. The *overt attention*, i.e. gaze movements, together with the scene as viewed from the subject's standpoint are recorded by using the WearCam system [12]. The order of the figures from left to right corresponds to the progress of the task. The cross superposed on the video corresponds to an estimated gaze position. It can be seen that the gaze is tightly bound to an object that is relevant to spatio-temporal requirements of the task. In spite of the presence of salient distractors, the gaze remains tightly locked on the current object of interest. This behavior cannot be predicted by the feature-based saliency maps, even with top-down extensions because in manipulation tasks perceptual processing is biased toward manipulation-relevant regions of the visual field, not toward the most textured or distinctively colored stimulus. The presented experiment is inspired by the experiments performed by Land et al. [13] and Hayhoe et al. [7] and it draws similar conclusions regarding the predominance of the influence of the motor system over the influence of low-level features on visual attention in a manual task. For more papers with similar insights obtained from human walking studies, see [14], [15].

guided visual search by an additional weighting of the feature channels with a top-down bias that comes from the prior knowledge about objects [22], [23]. One of the most influential top-down models of visual attention is the biologically-inspired model by Tsotsos et al. based on optimizing the visual search by using selective tuning of top-down, pyramidal, hierarchically-organized winner-take-all mechanisms [24]. The model addresses the problems of selection in an image, routing of information through the processing hierarchy and task-specific biases for visual attention. In their attentional model of human visual object detection, Oliva et al. have included the influence of the top-down prior information from visual context (i.e. where specific objects of interest should be located in the global scene configuration) on the saliency of spatial regions in natural scenes [25]. When the visual target is known beforehand, the top-down bias based on the similarity measure between the target and scene regions, where the features are orientation and spatial frequency histograms, can very efficiently guide visual search [26]. Interestingly, the model could accurately predict the distribution of human gaze fixations recorded while observing the same real-world scenes without having any sub-mechanism corresponding to bottom-up saliency activations.

B. Attention for robot vision

Related work in robotics is heavily influenced by the aforementioned Itti-Koch computational model of attention. Whereas most of the computational models implicitly assume covert attention shifts, i.e. no movements of the head and the eyes are involved, most robots are equipped with an active camera system, which makes them suitable for active, overt visual exploration. These robotic applications inherently rely on a saliency map-based scheme to evaluate visual stimuli, and then, instead of shifting covert focus of attention, they actively initiate saccadic movements of the cameras to bring the fixation to the most salient point in the visual field [27]. A number of robotic applications are primarily concerned with implementing saliency maps in order to achieve biologically-inspired saccadic and smooth-pursuit eye movements either with a single pan-tilt camera or a complete robot head [28]. These schemes have been extended to biologically inspired log-polar vision [29], [30]. Saliency-based attention has been studied in conjunction with exploration, development and learning for humanoid robots [30]. Attentional-based vision has been addressed as an aid to sociable robots to improve human-robot interaction [31], [32] and in imitation learning [33], [34].

C. Current shortcomings of attention-based models for robot vision and their biological solutions

Although the efforts made in the robotic community have been very fruitful, expanding theoretical foundations and providing practical applications of attentional mechanisms, the most prominent use of attentional schemes still remains applied to object tracking, scene exploration, mimicking the human visual system for robotic studies of development and for providing human-like visual behavior for sociable robots [27]. A very significant drawback of attentional models based on early perceptual saliency, for the purposes of visually driven motor control, is that an attentional relevance is computed solely on the structure determined from low-level visual stimuli projected on the retina, whereas neither the 3D structure of the environment, physical constraints such as body kinematics nor motor action plans are taken into account. The use of attention for active, real-time vision-based manipulation that relies on reliable visual information at each cycle of the control loop continues to be very limited. This is an issue we aim to address in this work. In particular, we identify the following three issues as critical: i) speed of computation, ii) distribution of focus of attention and iii) attentional relevance.

1) Speed of computation: Attention in primates evolved as a cheap, efficient and inherently embedded mechanism to select a small subset of abundant visual information for further, high-level processing. The primary reason for this

is to efficiently optimize the use of scarce computational resources. However, as previously mentioned, most work in robotics related to attention is motivated by the saliency model of Itti and Koch [21]. Regardless of the massively parallel architecture, constructing a saliency map is an extremely intense computational task. The best reported times on CPU-based implementations, highly specialized for efficiency, are of an order of ~50 ms for a single map [35], the time which doubles for a stereo system, after which, in addition, some high-level visual processing is done in the later stages in the visual processing pipeline. This prohibits applications of the classical saliency map approaches for fast real-world robotic problems such as real-time adaptation to perturbations in grasping tasks with obstacle avoidance.

2) Distribution of focus of attention: The majority of models of attention assume that a focus of attention, the so-called attentional spotlight, is a circular shaped region of a fixed radius [36], which is centered at a point with the highest saliency in the visual field. Zoom-lens models extend the attentional spotlight concept by allowing the radius of an attentional "window" to change with respect to task demands [37]. Both the spotlight and zoom-lens models restrict applicability of attentional mechanisms for real-world robotic scenarios in complex tasks because only one location in the visual field is (covertly) selected as the focus of attention, toward which the further attentional interest is oriented (covertly or overtly). A number of recent studies from visual neuroscience and psychology suggest that covert attention can take on a complex spatial arrangement [38]. Baldauf et al. have found that covert attention supports pre-planning of a rapid sequence of movements toward multiple reaching goals, by distributing peaks of attention along an intended reaching path [8], [9]. These findings show that covert attention can be distributed not only at one location, as overt attention, but rather simultaneously forms a complex "attentional landscape" in the visual field. Schiegg et al. found that covert attention can be split into multiple foci that are deployed in a way to pinpoint individual locations of intended contact points of the fingers during precision grasping [39]. The experiments with non-human primates have shown that visual receptive fields can even adapt after several minutes of the tool use by elongating their shape to covertly overlay the tool held in the hand [40], [41].

3) Attentional relevance: Computational models of attention have shown good performances and significant statistical similarity to human strategies in simple scene viewing and in guided search tasks [21], [20], but describing human gaze behavior in more complex tasks is far beyond their capabilities. The weakness of the majority of attentional models is that the top-down influences coming from the motor information are not taken into account. In a more recent work, Rodriguez-Sanchez et al. have developed an attentional model under the selective tuning framework that is able to recognize and attend complex motion patterns in the visual field [42]. Their model, which mimics some well-known properties of the mammalian visual stream, is able to detect and classify moving objects in a presented video stream, showing the behavior observed in real-world human psychophysical visual search experiments. Going along the line of emphasizing the importance of the motor information for visual attention, we hypothesize the observed mismatch between the predictions of the standard, low-level, feature-based attentional models and the actual deployment of attention in reach-to-grasp tasks (Fig. 2) is attributable to the fact that only low-level image features are taken into account by the models that compute attentional relevance, whereas the strong top-down bias from the arm motor system in reach-to-grasp tasks is completely ignored.

This is rather surprising, considering that there are numerous evidences that report on the very significant coupling between the motor system and attention allocation. Even in pure perceptual tasks, where vision does not support ongoing arm movements, the peripersonal space[1] receives a prioritized covert visual processing compared to the extrapersonal space [44], [40], [45], with the peaks of the attentional relevance of visual stimuli close to the hands [46], [47], [48], [49]. The importance of visual specialization of the peripersonal space is even observed at the level of the parts of the central nervous system. Neurophysiological studies in humans and non-human primates have revealed specialized circuits in the putamen, parietal cortex and ventral premotor cortex that are devoted to processing of visual stimuli within the peripersonal space [50], [51], [52], [40], [46]. Previc, in his well-known theory of visual field specialization, hypothesized that the visual prioritization of the peripersonal space emerges from functional relationships between the vision and motor systems [43]. In this view, the peripersonal space is inherently more visually salient than the extrapersonal space because it supports motor activities with the hands.

In studies that used overt gaze movements as a measure of attention and where the gaze is used to support physical actions, researchers have found that the gaze is driven by spatio-temporal task demands in simple navigation tasks [53], by manipulation in natural, well-known tasks [7], in moderately complex tasks involving obstacle avoidance [54], and in very complex tasks such as ball sports [55]. Similarly, studies that analyzed the distribution of covert attention in visuomotor tasks have shown similar results. Covert attention is brought to objects relevant to manipulation, even when reaching for multiple targets in a sequence [8], or in parallel by engaging bimanual manipulation [9]. The starting position of the hand [56] and its goal position [11] receive prioritized visual processing when preparing arm movements. Deubel and Schneider found that deployment of covert visual attention at an obstacle occurs when the obstacle obstructs intended arm movements, however, in cases when it does not obstruct intended manipulation it is not covertly attended [57]. Deployment of covert attention could be modulated by motor plans as tightly as to support planned finger movements during grasping [39].

Very few, if none, of the mechanisms reviewed in this subsection are utilized in the modern computational attentional

[1]The peripersonal space is defined as the space around the body within which an agent (a human, monkey or a robot) can manipulate objects without using locomotion to move the body, whereas the extrapersonal space is postulated as the space beyond the peripersonal space and its representation is used for navigation and orienting, see [43] for more.

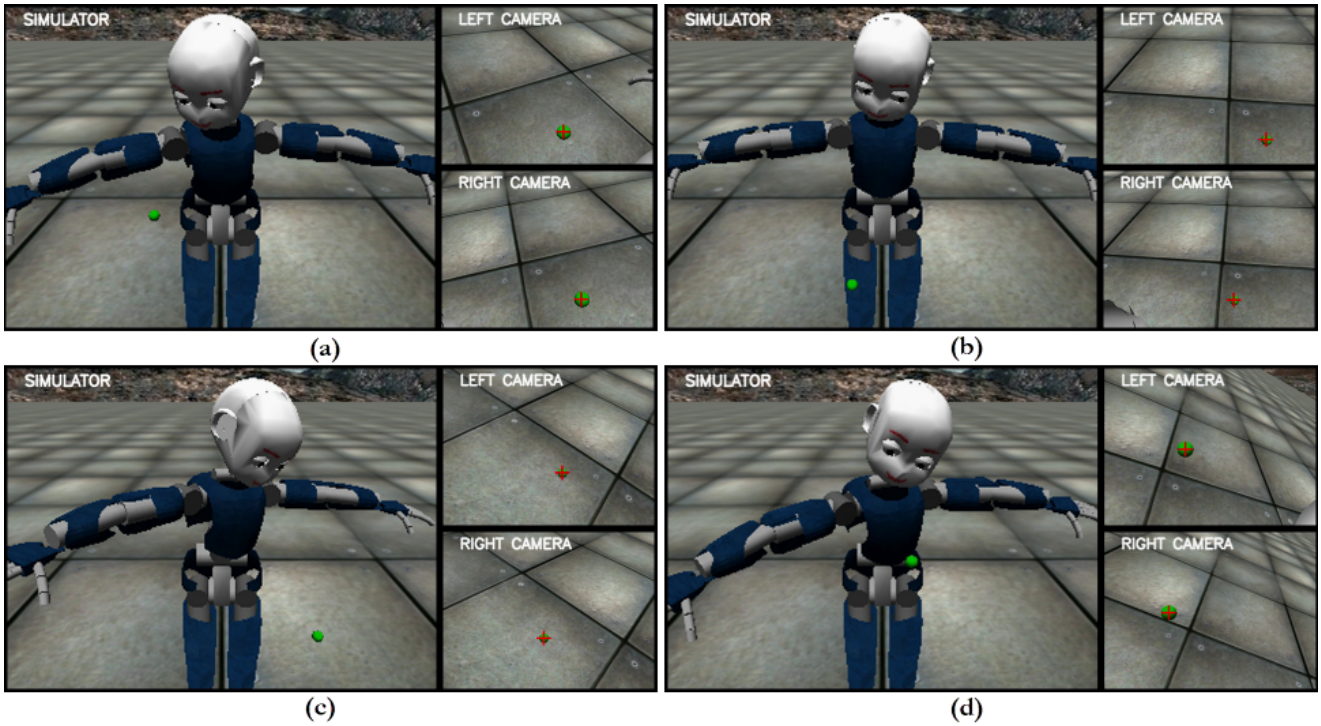


Figure 3. Exploratory behavior used to learn an adaptable model of the visuomotor transformation. The snapshots from the simulator (a-d) show several examples of exploratory configurations. The torso-neck-head-eye-joints (9 DoF) are sampled from the uniform distribution within their respective joint limits, and, similarly, the position of the green ball is sampled from the uniform distribution defined within the reachable space. For each sampled configuration, the encoders are read and the locations of the segmented ball in the stereo images are obtained. After the exploration, these data points are utilized to learn a neural network model of the workspace to the stereo image projections. The advantage of having such a model is that the model can be easily adapted with data points obtained from the real robot by taking similar exploratory procedure, in order to adapt the model to the discrepancies between the mathematical model and the kinematics of the real robot.

methods embedded in robotic visually-driven reaching and grasping. Taken together, biological studies indicate a clear dependence and an active modulation of visual attention on motor information. All these results suggest that low-level feature-based saliency is suppressed when an actor is engaged in visually-aided physical tasks, regardless whether the task is manipulation or navigation, whether the interaction with the object is performed in a parallel or in a sequential manner, and regardless whether gaze movements are suppressed or not. In plain words, in physical tasks, motor-relevant parts of the visual field are visually salient.

The aforementioned behaviors observed in these studies are elegantly explained and unified by the premotor theory of attention proposed by Rizzolatti and coauthors [58]. This theory argues that visual attention is a feature that emerges from the motor neural circuits that generate actions, i.e., cortical structures that are involved in arm movements are also responsible for constructing covert visual attention that accompanies the movements. In developing our model, we take the exact approach as argued by the premotor theory of attention: the attentional landscape is primed by the motor system. By equalizing motor-relevant as attention-salient, we aim at tackling the reviewed current weaknesses in the existing attention models. We demonstrate in this paper that motor-primed visual attention is a very efficient mechanism. We proceed further with the section that describes how the peripersonal space-primed attention and motor plans-primed attention landscapes are computed.

III. Peripersonal space-primed and motor plans-primed attention

We here first describe a method to compute projections from the workspace to the image plane. Once this transformation is obtained, it is used to construct two variants of visual attentional landscapes: (a) peripersonal space attention is based on the idea that visual attention is biased toward the reachable space of a robot, and (b) motor plans attention is a concept where attention is dynamically bound to motor plans of the robot.

A. The workspace to the image plane projection

In order to distribute visual attention with respect to both the peripersonal space and motor plans of a robot, we first need to obtain a transformation that will map the points from the spatial coordinates to the image planes.

1) Projections to the image plane: Let the Cartesian workspace position be represented as $x \in \mathbb{R}^3$, and the kinematic configuration at the current time of the torso-neck-arm represented with the torso, neck, and head joints as $q \in \mathbb{R}^9$, the transformation function of the form:

$$p_i = f_i(c), \tag{1}$$

where $c \in \mathbb{R}^{12}, c = \begin{bmatrix} x \\ q \end{bmatrix}$, and $p_i \in \mathbb{R}^2$ represents the projection of the trajectory, taking into account the kinematics of the torso, neck and eye, to the image plane of the i-th camera, where $i = \{left, right\}$.

A classical, straightforward approach would be to compute a sequence of kinematic transformations through the torso-neck-head kinematic chain in order to obtain the extrinsic camera parameters, and use them together with the intrinsic parameters of the camera to obtain the projective transformation. For a stationary camera, calibration of all the camera parameters can be easily accomplished by formulating the problem as linear regression and solving it by using the least-squares approach. However, for cameras mounted on a moving robot's head, the problem includes the torso-neck-head joints. This imposes the need for calibration of the kinematic chain because most often a real robot differs from its ideal kinematic model. Hence, the linear problem of calibration for a static camera becomes highly nonlinear for a camera mounted on the head as we include the torso-neck-head joints as independent variables.

Clearly, an alternative solution is to rely on a non-linear approximation using any of the standard machine learning techniques for non-linear regression. The robot would explore in a babbling-like manner a set of kinematic configurations, and during this exploration it would segment an object (e.g., a small colored ball) placed at a randomly chosen position from a set of known positions in the workspace. The data obtained during the exploration (encoder readings of the joints in the torso-neck-head chain, the position of the object in the workspace and its projection to the camera planes) would be used to learn a mapping function. A problem associated with this approach is that the babbling-like exploration with the real robot is very costly because in order to build a reliable estimate of this nonlinear mapping, the size of a training set needs to be arbitrarily large to be representative, usually of an order a few thousand data samples.

Here we take an intermediate step that represents a compromise between the two previously described approaches. The idea is to take advantage of the simulator of a robot in order to obtain a large number of training samples by employing babbling, and use this data set to estimate an initial set of parameters for the mapping model (Fig. 3). Once a neural network workspace to the image plane projection is learned in the simulator, we incrementally adapt the model with a fraction of the data ($\sim$100 data points) from the real robot in an incremental manner. The exploratory procedure with the real robot is identical to the one performed in the simulator (Fig. 3). The only difference with the real robot is that the object to be segmented, instead of the green ball programmed to be moved in the simulator, is a green-colored marker placed on the palm of the robot. The robot in an exploratory fashion moves the arm and the torso-neck-head system, the 3D position of the marker is read from the forward kinematics of the arm and the marker image projections are obtained after a color-based image segmentation procedure.

2) Neural network approach: A feed-forward neural network is a suitable machine learning algorithm for our application for several reasons [59]. Feed-forward neural networks can compute multi-input-multi-output functions. Their output is very fast to compute in real-time because the computation consists of a short sequence of matrix-vector multiplications, followed by (non)linear transfer functions. Feed-forward neural networks are suitable for incremental learning, either in the batch or in the stochastic, online mode. This allows us to first estimate this function from the data in the simulator, and then adapt it with the data from the real robot.

The parameters of an architecture of neural networks for transformation from the workspace to the image coordinates (i.e. number of layers and the number of hidden units, etc.) are determined by using grid-search on the mean squared error (MSE) between the recorded image projections and retrieved projections from the model. We tested 10 different network architectures, and for each architecture, we performed 10 learning runs in order to ensure robustness with respect to random initialization of network parameters. We used the Levenberg–Marquardt optimization algorithm with early-stopping in order to prevent overfitting [59]. The recorded data set is randomly partitioned for 70 % of the data devoted to training, 15 % data for validation and 15 % data for testing. The lowest MSE on the testing set is obtained using two hidden layers with 25 nodes in each hidden layer. Transfer functions in the hidden layer are hyperbolic tangent sigmoid, and in the output layer are linear. The data set is normalized to obtain zero mean and unity variance. In order to get the real-time performances, a network class is implemented in C++ by using linear algebra functions from OpenCV library [60]. The time needed to transform 50 points by using neural nets to the image planes of both cameras is less than 1 ms.

B. Peripersonal space-primed attention

In order to be able to distribute visual attention with respect to the peripersonal space of a robot, (a) we need to have a transformation that will map the peripersonal space to the image planes (as described in the previous section III-A), and (b) we need to obtain a representation of the peripersonal space that we will map to the stereo images as the robot takes different postures. We next proceed with describing how we obtain the representation of the peripersonal space and how we learn the peripersonal space-primed attentional landscape.

1) Representation of the peripersonal space: For reachable space modeling, classical methods such as polynomial discriminants and geometric approaches compute the boundaries of the robot's reachable space (reviewed by Kim et al. [61]). The limitations of these methods are that they can only be applied to special kinematic chains and they model the boundary of the reachable space, without any notion regarding which locations of the reachable space are more likely to be attended. We take here an exploratory, sampling based approach that overcomes these two difficulties.

We model the peripersonal space by commanding the robot to explore reachable positions by randomly varying the arm joint angles. More specifically, we sample the joint values from the uniform distribution defined over the feasible joint ranges, and we read the achieved 3D end-effector positions from the robot's forward kinematics. Once this exploration is

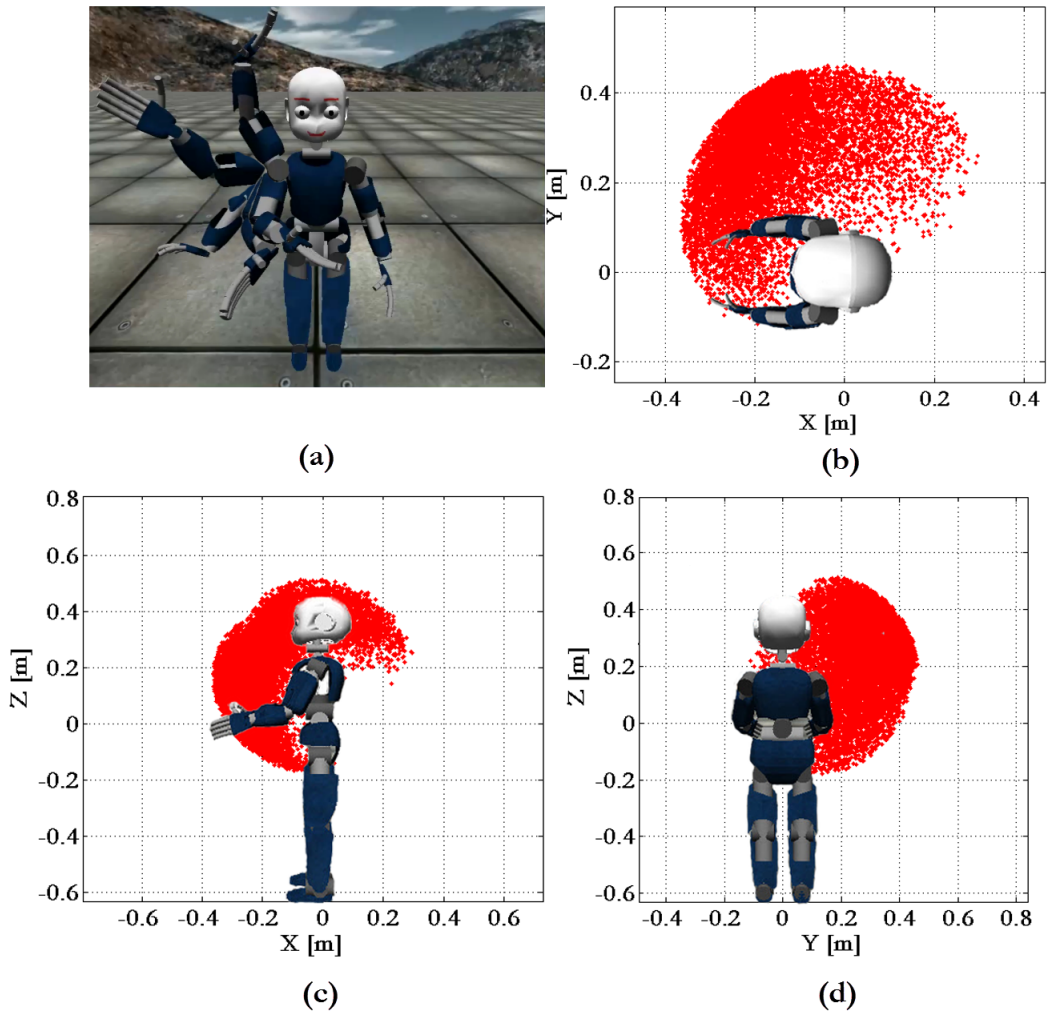


Figure 4. Exploratory behavior used to model of the peripersonal space for the right arm of the iCub robot. Figure (a) represents several exploratory movements captured in the simulator and superimposed. Figures (b-d) show the sampled cloud of data points with respect to the robot's body in XY (b), XZ (c) and YZ plane (d).

carried out, we store recorded reachable points in a database. Fig. 4 shows the exploratory procedure that we take and the obtained, sampled representation of the peripersonal space. The point cloud that models the sampled representation of the workspace of the robot obtained in the simulator is used without any adaptation to the real robot due to the fact that the differences between the workspace volumes of the real robot and the simulated one when projected to the image planes are negligible. One additional justification for this is that sampling the reachable space with the real robot would be a very expensive and slow process.

2) Attentional landscape: After we obtain the representation of the peripersonal space, we model the distribution of attention with respect to the peripersonal space. We sample the eye-neck-head joints from the uniform distribution within the joint limits and, for each sampled configuration, we project the previously sampled cloud of the reachable space points by using the previously learned mapping (presented in III-A) to the stereo images. This procedure is shown in Fig. 5.

The bubble-shaped cloud of the end-effector locations that models the peripersonal space (Fig. 4) projects as an ellipsoid-shaped scatter to the left and right cameras (Fig. 5). We use a bivariate Gaussian distribution to model the scatter of the projected points on the image planes, which represents a parametric representation of a peripersonal attentional landscape, as formulated:

$$\Lambda_{i,t}(p;\mu_{i,t},\Sigma_{i,t}) = \frac{1}{\sqrt{(2\pi)^4 \mid \Sigma_{i,t} \mid}} e^{-\frac{1}{2}((p-\mu_{i,t})^T(\Sigma_{i,t})^{-1}(p-\mu_{i,t}))}, \quad (2)$$

where t is the index of the currently sampled configurations and the corresponding projections, $i = \{left, right\}$, and $\mu_{i,t}$ and $\Sigma_{i,t}$ are the mean and the covariance matrix, respectively. In this case, the bivariate Gaussians, one for the right and one for the left image, are parametric representations of the peripersonal space attention for the stereo setup given the current neck-head-eye posture of a robot.

Before we proceed with learning of a function that maps the neck-head-eye posture to the parametric representation of the

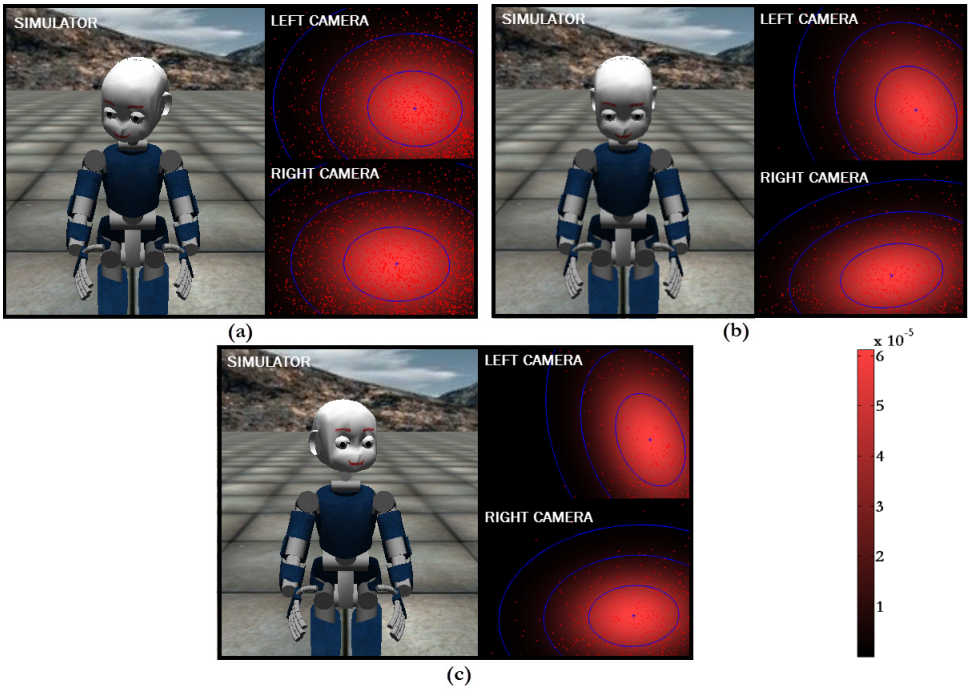


Figure 5. Exploratory behavior used to learn a model of peripersonal space-primed attention (a-c). The figures represent several exploratory movements. For each randomly generated neck-head-eye posture, we project the sampled set of reachable points to the stereo image planes. Using bivariate Gaussian distributions to model the elliptical envelopes of projections of the bubble-shaped cloud to the image planes is an intuitive choice. In this case, the bivariate Gaussians, one for the right and one for the left image, are parametric representations of the peripersonal space attention for the stereo setup given the current posture of a robot. The reddish heat maps correspond to the values of the density of the projections. Finally, a mapping from the neck-head-eye joint angles to the parametric representation of attention is learned by using these data. This mapping is used in the run-time to infer how peripersonal attention should be distributed given the set of the neck-head-eye joint angles.

attentional landscape (the mean and the covariance), we must take into account that the covariance matrix, inferred from such a mapping and used to compute the attentional landscape, must be symmetric and positive definite to ensure the validity of the Gaussian distribution. One solution is to enforce this by projecting the inferred covariance matrix (only symmetric but no guarantees of positive-definiteness) onto the set of symmetric positive definite matrices by using the constrained convex optimization programming. However, addressing this problem involves iterative optimization procedures, which we want to avoid for maximizing computational efficiency. Here we use an alternative approach. We first decompose the covariance matrix into the product of a lower triangular matrix $L_{i,t}$ and its transpose by using the Cholesky factorization:

$$\Sigma_{i,t} = L_{i,t}L_{i,t}^T,\ L_{i,t} = \begin{bmatrix} L_{1,i,t} & 0 \\ L_{2,i,t} & L_{3,i,t} \end{bmatrix}. \quad (3)$$

Next we proceed with learning a mapping $\lambda_{i,t} = g_i(q_t)$, defined from the current joint angles $q \in \mathbb{R}^6$ to the tuple $\lambda_{i,t} \in \mathbb{R}^5, \lambda_{i,t} = [\mu_{1,i,t}, \mu_{2,i,t}, L_{1,i,t}, L_{2,i,t}, L_{3,i,t}]^T$, which is an ordered, column-vector arrangement of the elements of $\mu_{i,t}$ and $L_{i,t}$. This mapping is learned with a feed-forward neural network, by using a similar procedure to the one explained in Section III-A. In the run-time, for a given configuration q^*, we infer $\tilde{\lambda}_{i,t}$, i.e.,$\tilde{\mu}_{i,t}, \tilde{L}_{i,t}$, from function g_i. We then compute the attentional landscape as follows:

$$\Lambda_{i,t}(p; \tilde{\mu}_{i,t}, \tilde{L}_{i,t}) = \frac{1}{C} e^{-\frac{1}{2}((p-\tilde{\mu}_{i,t})^T (\tilde{L}_{i,t}\tilde{L}_{i,t}^T)^{-1}(p-\tilde{\mu}_{i,t}))}, \quad (4)$$

where C is a normalization constant. The reconstructed covariance matrix, computed as the product $\tilde{L}_{i,t}\tilde{L}_{i,t}^T$, is a symmetric positive definite matrix. Considering that the Cholesky lower triangular matrix represents the measure of deviation from the isotropic Gaussian, we can constrain computation of the attentional landscape within the ellipse obtained by multiplying the unit circle $D = \{p \in \mathbb{R}^2 \mid \|p\|_2 = 1\}$ with $\tilde{L}_{i,t}$ and translating the product by $\tilde{\mu}_{i,t}$:

$$E_{i,t} = \sigma\tilde{L}_{i,t}D + \tilde{\mu}_{i,t}. \quad (5)$$

$\sigma \in \mathbb{R}$ is a free parameter that corresponds to the number of standard deviations at which one wants to compute the ellipse, and it is usually set at $\sigma = 3$. The value of the attentional landscape outside the 3σ-ellipsoid is insignificant to affect the distribution of attention and can be neglected. For this reason, we cut-off the attentional landscape at zero outside the 3σ-ellipsoid to avoid computing Eq. 4 at these pixels to gain computational efficiency.

C. Motor plans-primed attention

The peripersonal space primed attention could be seen as a general, multipurpose technique, to compute the distribution of attention to the image regions that correspond to the entire peripersonal space. It might or might not involve reaching and

grasping movements. However, because peripersonal space-primed attention is bound by the whole reachable space, it does not utilize particular motor plans of a robot. Additional constraining of the attentional landscape around motor plans-relevant regions results in additional computational savings and more localized visual processing. We here present a way to further constrain the attentional landscape, with respect to motor plans of the robot. This is a more specialized technique than peripersonal space-primed attention.

We first describe our robotic eye-arm-hand controller, developed in our previous work [62], [63], which we use to generate reaching and grasping movements and to forward-plan the arm-hand reaching trajectory. Learned eye-arm-hand Coupled Dynamical Systems (CDS) are used in order to "mentally simulate" the consequences of intended actions, more specifically, to compute (i.e. plan) an intended trajectory and to identify obstacles. This mentally simulated arm reaching trajectory is transformed to the image planes of the stereo cameras. The projected mentally-simulated trajectory is used to compute an attentional landscape.

1) Eye-arm-hand controller: The controller is based on the CDS framework, where the main idea is to estimate separate Autonomous Dynamical Systems (DS) that correspond to each body part (one DS for the eyes, one for the arm and one for the hand), and then couple them explicitly. This controller consists of five building "blocks": three dynamical systems and two coupling blocks between them. They are organized in the following order: eye dynamics $\rightarrow$ eye-arm coupling $\rightarrow$ arm dynamics $\rightarrow$ arm-hand coupling $\rightarrow$ hand dynamics, where the arrow direction indicates the direction of control signals. The gaze DS is the master to the arm DS, and the arm DS is the master to the hand DS. Fig. 6 schematically illustrates the architecture of the CDS. We used human motion capture data recorded in a reaching-and-grasping task to estimate the parameters of the controller [62], [63]. The time-invariant properties of the CDS allow rapid adaptation to spatial and temporal perturbations, where the explicit coupling between each dynamical system ensures that their behavior is correctly synchronized, even when the motion is abruptly perturbed far from the motion recorded in human demonstrations.

This framework allows us to perform visuomotor coordination in the presence of an obstacle, as well. We use the CDS mechanism in order to mentally simulate the consequences of planned arm movements, specifically to detect objects that obstruct the intended reach-for-grasp actions and to identify them as obstacles. The motion of the arm toward the target is calculated by integrating the dynamics of the CDS until each DS reaches its attractor. The arm end-effector is modeled as a point that moves along the estimated trajectory. Obstacle objects in the workspace are modeled as cylinders that enclose the actual dimensions of the object and also account and account for the fact that the hand was modeled as a point. By taking this approach, we are able to reliably detect collisions with the fingers in our forward planning scheme, even though the hand is modeled as a point, which results in a simplistic collision checking scheme. An obstacle is used as the intermediary target for the visuomotor system, which allows us decompose the obstacle avoidance task in two segments: from the start to the obstacle and from the obstacle to the target. In the first part of the task, the arm DS moves under the influence of the attractor placed at the via-point. The hand DS is driven by the attractor placed at the hand configuration when the palm reaches the closest point (along the trajectory computed ahead of time) to the obstacle. Coupling the hand motion with respect to the obstacle is advantageous because it modulates the shape of the hand such that collisions between the fingers and the obstacle are avoided during obstacle avoidance. The goal hand configuration for passing the obstacle at the closest distance is determined by observing the average hand configurations of our subjects in obstacle avoidance trials. The position of the via-point is determined with respect to the obstacle, such that its displacement vector from the obstacle position is oriented in either an anterior or posterior direction, for the length that corresponds to some safety distance between the centroid of the palm and the obstacle. We choose the direction of a displacement of the via-point (anterior or posterior) to correspond to a side of the obstacle where a collision is estimated to occur. This yields a minimum-effort obstacle avoidance strategy. In the second part of the task, after the obstacle is passed, the CDS is driven toward the object to be grasped. Predefining the safety distance at which the hand passes the obstacle is a biologically-motivated decision, motivated by a series of studies with humans. The arm end-effector passing through the via-point at the safety distance from the obstacle and hand adaptation, with respect to the obstacle, ensures that the hand will not collide with the obstacle. During obstacle avoidance, the primary modulation of the arm is controlled in Cartesian space, which, together with controlled hand preshape, ensures that the end-effector avoids the obstacle. In addition to this primary modulation, the secondary modulation consists in suggesting the desired arm joint postures suitable for obstacle avoidance. Favorable joint configurations are first learned from human demonstrations, and after that, they are in the run-time inferred and provided to the IK solver. We utilize the mentally simulated palm trajectory obtained by integrating the CDS in order to bias visual resources to motor-relevant parts of the visual field, which we describe in the next section. For more about the CDS framework, its extension to obstacle avoidance and the biological motivations of the visuomotor controller, see Lukic et al. [63].

2) Attentional landscape: The mentally-simulated trajectory of the arm, from the current position to the final position at the current time t, is represented as $x_t^n \in \mathbb{R}^3$, $\forall n \in [1, N_t]$, where N_t represents the total number of discrete samples. This mentally-simulated trajectory at every cycle of the control loop t is obtained from the CDS controller, explained in Section III-C1. The kinematic configuration at the current time t of the torso-neck-head is represented with the torso, neck, and head joints $q_t \in \mathbb{R}^9$, $\forall t$. We use the previously learned transformation function, presented in Section III-A, to perform this mapping:

$$p_{i,t}^n = f_i(c_t^n),\ \forall n \in [1, N_t]\,, \tag{6}$$

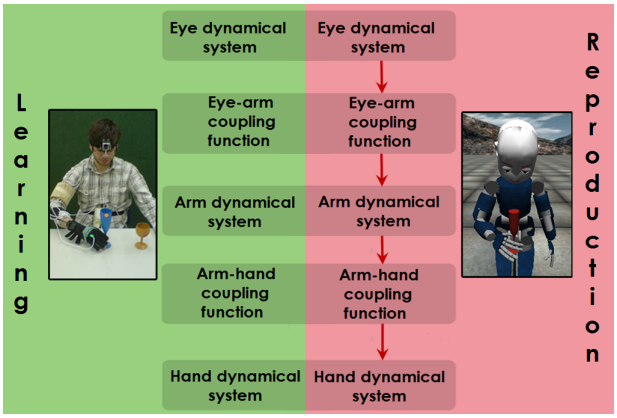


Figure 6. CDS-based eye-arm-hand controller. Left (green) part of the figure illustrates how the CDS model is learned. Reproduction of motion on the robot is indicated on the right side of the figure (red part). CDS includes five building blocks: three dynamical systems (the eyes, the arm and the hand) and two coupling models: eye-arm coupling and arm-hand coupling.

where $c_t^n \in \mathbb{R}^{12}$, $c_t^n = \begin{bmatrix} x_t^n \\ q_t \end{bmatrix}$, and $p_{i,t}^n \in \mathbb{R}^2$ represents the projection of the trajectory to the image plane of the i-th camera, where $i = \{left, right\}$.

After we project the mentally-simulated trajectory to the image planes, we construct an attentional landscape which associates high attentional relevance close to the mentally-simulated trajectory perceived in the image coordinates. To compute the attentional landscape, i.e. a measure of visual processing priority, we use a bivariate kernel smoothing function, where kernels are placed at every point of the projection of the mentally-simulated trajectory to the image planes. Formally, we compute an attentional landscape for each camera i as follows:

$$\Lambda_{i,t}(p) = \frac{1}{N_t h^h h^v} \sum_{n=1}^{N_t} K(p - p_{i,t}^n), \tag{7}$$

where $p \in \mathbb{R}^2$ corresponds to two-dimensional pixel coordinates of the image plane,

$$K(p - p_{i,t}^n) = k\left(\frac{p^h - p_{i,t}^{n,h}}{h^h}\right) k\left(\frac{p^v - p_{i,t}^{n,v}}{h^v}\right), \tag{8}$$

where $k(.)$ represents a kernel and h^h and h^v are kernel widths along the horizontal and vertical image dimensions. Kernel widths along the horizontal and vertical image dimensions are the parameters of the algorithm that are hand-tuned. The greater value of the kernel width, the larger area of the image is to be processed in the corresponding direction and, hence, additional computational costs become associated with it. A good rule of thumb is to select greater kernel widths when spatial perturbations of objects in the scene are expected, to better cover the scene to detect target perturbations and possible sudden obstacles entering the workspace. We tested both Gaussian kernels and triangular kernels, and we choose to use triangular kernels because they are faster to calculate. The triangular kernel is expressed as follows:

$$k(z) = \begin{cases} 1 - \left|z\right|, & \left|z\right| \leq 1 \\ 0 & otherwise \end{cases}. \tag{9}$$

The kernel smoothing function assigns high values of attentional relevance close to the mentally-simulated trajectory projected to the image planes of stereo cameras, which decrease in the directions away from the trajectory (Fig. 1). The attentional landscape is used to guide image processing in order to efficiently distribute limited visual resources. The part of the image with higher attentional relevance draws more visual processing, and the opposite is true. In the next section, we explain how we distribute visual processing with respect to the visual attentional landscape, both peripersonal space-primed and motor plans-primed.

IV. Visual Processing Prioritizing Attentional Relevance

In Section III, we presented two types of attentional landscapes that can be utilized to distribute visual attention emerging from the motor system. In order to detect objects relevant to the task at hand, a robot must process stereo images. In this section, we propose two ways to use the attentional landscape to guide visual processing. These two techniques make our approach general enough to be used as a pre-modulating technique to almost any kind of standard image processing detectors and segmentation techniques (pixel-by-pixel color segmentation, histogram-based detectors, Viola-Jones, SIFT, SURF, etc.). The two processing schemes that will be presented apply to both peripersonal space-primed and motor plans-primed attention.

A. Thresholding

One simple approach suitable for pixel-by-pixel color processing and interest point detectors-descriptor approaches is to distribute visual processing to the region of the image where an attentional landscape $\Lambda_{i,t}(p)$ is higher or equal than some threshold d_i. It is easy to empirically estimate the computational time for processing the entire image and from this value estimate cost per pixel. By sorting pixels with respect to ascending values of their attentional relevance, we can pick a number of pixels corresponding to the available computational resources. From this sorted array, we can easily compute the threshold d_i on the attentional landscape. An approximate value of the threshold can be determined in ~3 ms for 4800 subsampled pixels by using the Quick Sort algorithm.

B. Sampling

A number of image processing techniques employ image processing within a scanning window, e.g. Viola-Jones detector, histogram-based detector, Rowley-Baluja-Kanade detector, etc. Here the task is to determine the position of the scanning windows with respect to an attentional landscape $\Lambda_{i,t}(p)$, in order to have more dense scanning where the attentional relevance is large, and less dense scanning in spatial regions with low attentional relevance. Because we use either a kernel smoothing function or a Gaussian function to build an

attentional landscape, we can treat the attentional landscape as a bivariate probability density function and use any kind of sampling techniques to sample spatial locations of scanning windows. Again, we can empirically obtain a cost associated to process the image in each window, and from the total visual resources, calculate the number of points to sample from the attentional landscape. We use the Gibbs sampling method [64]. We choose the Gibbs sampling instead of other sampling procedures such as the general Metropolis-Hastings algorithm[2] because the acceptance rate of sampled proposed values is 1[3], which makes it a very efficient procedure. The procedure operates as follows:

1) start with an initial pixel location: $p_{i,0} = [h_{i,0}, v_{i,0}]^T$;
2) for $j = 1, 2, \ldots, M$;
3) sample $h_{i,j}$ from the conditional distribution $\Lambda_{i,t}(h \mid v_{i,j-1})$ by using the inverse transform sampling;
4) sample $v_{i,j}$ from the conditional distribution: $\Lambda_{i,t}(v \mid h_{i,j})$ by using the inverse transform sampling;
5) store $p_{i,j} = [h_{i,j}, v_{i,j}]^T$, increment j and loop over steps 3-5 for the given number M of scanning windows;
6) return the set of sampled points: $P = \{p_{i,1}, \ldots, p_{i,M}\}$ (locations of scanning windows).

The Gibbs sampler and inverse transform sampling function embedded in it are implemented with look-up tables as C-arrays for efficiency. The time for querying the Gibbs sampler is ~3 ms for an attention landscape of size 320×240 for 50 sampled scanning windows.

1) Adjustment when sampling from the peripersonal space-primed attentional landscape: In Section III-B, we presented the method for modeling the peripersonal space attention with one bivariate Gaussian per stereo image. The bivariate Gaussian is suitable for modeling the projection of the 3D peripersonal space blob to the image plane, as we illustrate in Fig. 5. Once this representation is obtained, it is used to perform image processing according to it. For processing by using the thresholding-based approach, this representation of the attentional landscape can be directly used, however, for the sampling-based approach, we find that it is better to slightly balance it. The steeply rising profile of the Gaussian distribution biases sampling toward its centroid. When we sample a smaller number of windows, this could lead to the case that the objects that lie closer to the boundary of the reachable space are missed. For this reason, we propose using a balanced version of the peripersonal space-primed attention (Section III-B) when doing sampling-based image processing. A balanced peripersonal space-primed attentional landscape is defined in a form of a mixture between the obtained bivariate Gaussian (Eq. 4) and the uniform distribution $U(p)$:

$$\Lambda_{i,t}(p; \tilde{\mu}_{i,t}, \tilde{L}_{i,t}) = \pi \frac{1}{C} e^{-\frac{1}{2}((p-\tilde{\mu}_{i,t})^T (\tilde{L}_{i,t}\tilde{L}_{i,t}^T)^{-1}(p-\tilde{\mu}_{i,t}))} + (1-\pi)U(p), \; U(p) = \begin{cases} c, & c \in domain \\ 0 & otherwise \end{cases}, \quad (10)$$

where $\pi \in [0, 1]$ is the mixing probability, which is a parameter that can be hand-tuned according to desired behaviors and $domain$ refers to the area in the image for which $\Lambda_{i,t}$ is selected to be computed, either for the entire image or 3σ-ellipsoid. Creating the mixture between the Gaussian and the uniform distribution flattens the original Gaussian profile, which results in more spread out sampling and, hence, better coverage of image regions that correspond to the spatial regions lying closer to the boundaries of the peripersonal space. Again, we choose to constrain computations within the 3σ-ellipsoid $E_{i,t}$.

C. Closing the loop: from covert attentional landscape to overt eye movements and manipulation

It is noteworthy to mention that we recompute and sample the attentional landscapes maps at every cycle. This implies that there is no requirement to implement the IOR mechanism and deal with the problems with the change of coordinates associated with standard saliency models [27], which simplifies our approach and hence reduces the overall computational time.

As described in the previous section, when the attentional landscape is constructed, the top-down visual scan is performed in the spatial regions that have high relevance. These two stages correspond to covert visual attention. In the case of motor-primed attention, after the targets (and/or obstacles) are detected, the overt gaze movements are initiated toward the first intermediary target in a synchronous manner together with the arm and the hand motion by using our CDS eye-arm-hand controller [62], [63]. In a no-obstacle task, the eye-arm-hand system is directly driven toward the target. In tasks with obstacle avoidance the eye-arm-hand system is driven toward the obstacle, which is treated as an intermediary target for the visuomotor system, as explained in Section III-C1. When the obstacle is avoided, the system is driven toward the object to be grasped.

V. Experimental validation

We validate our method in the iCub simulator and with the real robot with a task of visual exploration for initial object detection (peripersonal space-primed attention), and reaching for and grasping a kitchenware object (motor plans-primed attention). Resolution of the stereo cameras in the setup is 320×240. We verify this approach with two well-known standard image processing techniques. For the first visual detector, we select a scanning window hue-saturation histogram-based detector. We implement this detector by using functions from the OpenCV library [60]. For the second detector, we selected

[2]The Gibbs sampling algorithm can be viewed a special case of the Metropolis-Hastings algorithm, which belongs to a wider class of Markov Chain Monte Carlo (MCMC) methods. The basic approach adopted in MCMC methods is to draw correlated data points from the obtained probability distribution based on the constructed Markov chain on the state space that has the target distribution as its equilibrium distribution. For more, see [64].

[3]At each iteration, the Metropolis-Hastings algorithm picks a candidate for the next sampled data point based on the currently sampled data point. Then, with some probability of acceptance (i.e. acceptance rate), the candidate point is either accepted or rejected, to ensure that the fraction of time spent in each visited state is proportional to the target density. In the case of the Gibbs sampling, the acceptance rate is always 1, meaning that all proposed candidate data points are kept. For more, see [64].

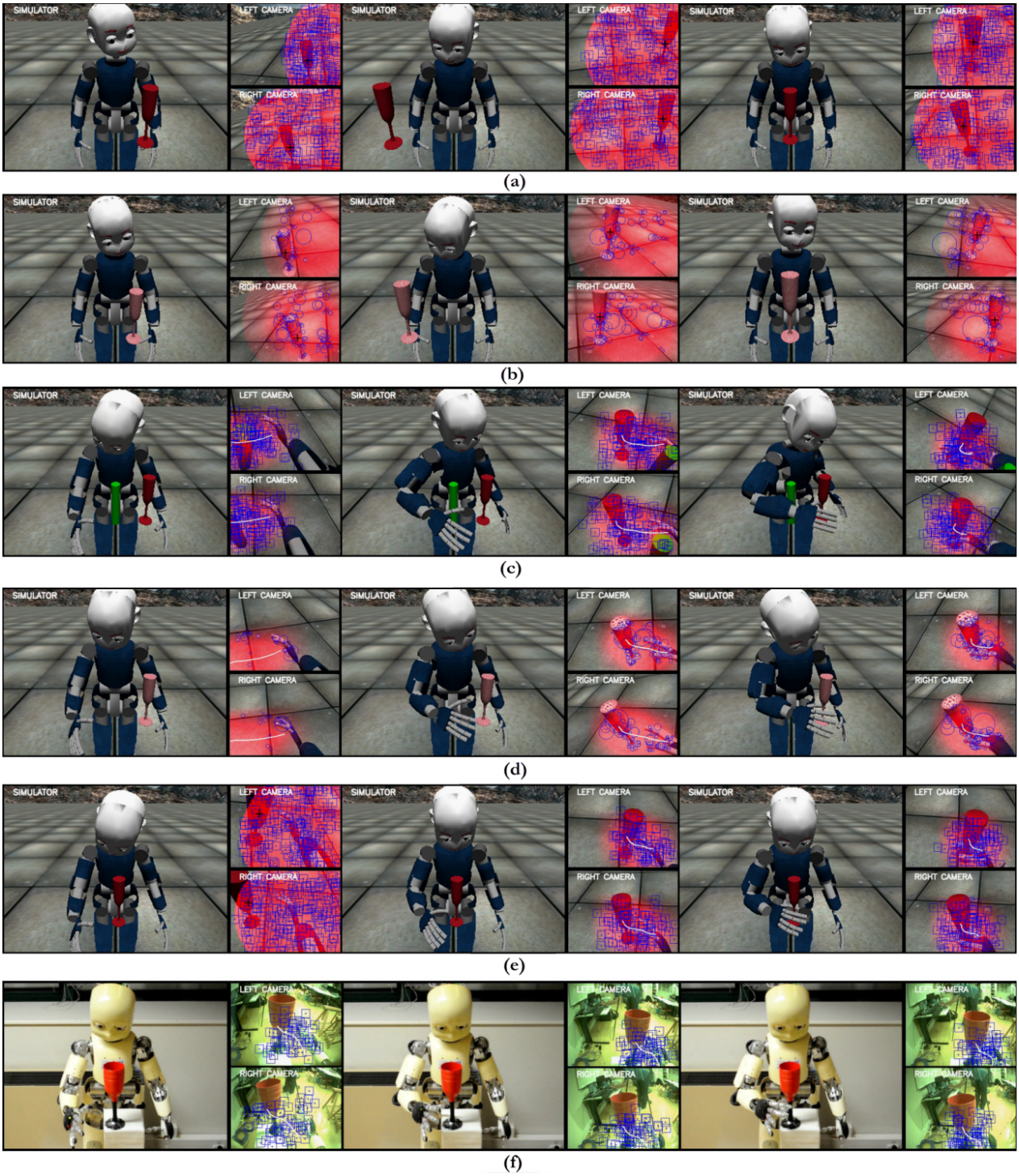


Figure 7. Experiments of visual exploration for object detection (a-b) and visually-guided reaching and grasping in the iCub's simulator (c-e), in two different scenarios with two detectors, and the real robot (f). The reddish blend shows the superimposed attentional landscape used to drive visual processing (for the peripersonal space-primed attention with the histogram-based detector (a) we are sampling from a modified version, computed as in Eg. 10 with $\pi = 0.2$). The figures (a) and (b) represent snapshots from the experiments where visual processing is prioritized to the peripersonal space (peripersonal space-primed attention), for histogram-based detector and SURF, respectively. The blue squares are scanning image windows for which visual features are computed. The robot adopts a random configuration and the object adopts a random position within the reachable space. Figures (c-f) show the context of motor plans-primed attention, namely, the execution of eye-arm-hand coordination from the start of the task (left) until successful grasp completion (right). The white line corresponds to a mentally-simulated arm trajectory that is projected to the image planes of stereo cameras. Figure (b) corresponds to the obstacle scenario with histogram-based detector. Figure (c) corresponds to the obstacle scenario with histogram-based detector. Figure (d) corresponds to the no-obstacle scenario with SURF detector. The blue circles correspond to detected strong feature points. Figure (e) shows how a combination of both approaches: the peripersonal space-primed attention is used to bootstrap initialization of the motor plans-primed attention. The bottom row (f) corresponds to the no-obstacle scenario with histogram-based detector with the real robot.

SURF [65][4]. SURF is a powerful detector because it provides visual features that are robust to moderate changes of the perspective. Because it computes feature point descriptors, it provides the ability to detect partially occluded objects. However, SURF (together with a family of similar detectors like SIFT, GLOH, etc.) is very computationally demanding, with the cost being double for a binocular system, hence it has limited applicability for manipulation where the stereo vision is used in the loop. The total time to process a stereo pair of images in the standard, full-blown way, is for the histogram based detector with the window size 20×20 is 168 ms and for SURF with the Hessian threshold set to 300 is 515.5 ms.

We first test both detectors in the context of peripersonal space-primed attention. The time needed to infer the parametric representation of the attentional landscape by using feed-forward neural networks is negligibly small, close to a tenth of a millisecond. Computing the peripersonal attentional landscape image requires 35.5 ms. These computations are common for both image processing techniques. Sampling from the relevance images, for the histogram-based detector, requires 7 ms for the stereo setup for 50 image windows per image. Performing sparse image processing for these windows takes 26.5 ms. These times sum up to 69 ms for the peripersonal-space histogram based visual detection. We can see that with our approach we can save 99 ms for each pass through the control loop (speed up factor ~2.4 ×). For SURF, thresholding takes 6.5 ms and processing 30 % of the image pixels with the highest salience takes 280 ms, which sums up to the total time of 322 ms for our approach. We can see that this saves 193.5 ms per pass (~1.6 × faster). Figures 7(a-b) show the simulated results.

For motor plans-primed attention, we use the similar approach, the only difference is that this, more specialized visual attention, is used to aid the ongoing movements. For both detectors, the common computations involve a projection of the mentally-simulated trajectory to the image plane and computing a motor-primed attentional landscape. The cumulative time for calculating a projection of the forward-planned trajectory to the image planes and computing attentional image landscapes is 19 ms (1 ms for projection and 18 ms for computation of the landscapes). For the histogram-based detector, sampling time for 50 windows is the same as in the peripersonal version, 7 ms, and similarly, the image processing time is 28 ms. The overall time for motor plans-primed histogram-based image processing is only 54 ms, i.e. ~3.1 ×(114 ms) faster than the naive image processing with a uniformly sliding window. For motor-plans primed attention with SURF, again, thresholding requires 6.5 ms and processing 30 % of the image pixels of the most relevant pixels takes 281 ms. The total time for our approach with SURF is 306.5 ms, which is ~1.7 ×faster than the classical, full-blown image processing. Figures 7(c-d) show the scenarios. Figure 7(f) presents the experiments with the motor plans-primed attention and the histogram-based detector with the real iCub robot.

The presented schemes could be used independently of each other, as previously discussed and as shown here, however, they could work even better if used together. In order to plan the movements for actions (for estimation of future movements and for updating the visual scene by using visual processing driven by motor plans-primed attention), a robot must have some initial guess where the object might be. Of course, to initialize the procedure one could scan the entire images first and then in the further iterations apply reduced processing by utilizing the motor attention and updating the knowledge about the object state from the vision system. However, for this initial exploration, we could use the peripersonal space attention to constrain the initial visual search. Once the robot starts to move, it switches to the motor plans-primed mechanism. Figure 7(e) shows how these two attentional mechanisms work together.

Clearly, the presented experiments show that if we choose to intelligently process the images, prioritizing valuable image resources to motor relevant plans of the images, we can speed up visual computations by up to a factor of 3 times compared to standard uniform image processing approach, where all pixels have the same priority and hence they are processed accordingly, without any discrimination what is motor relevant from what is not.

Finally, it is important to mention that, in addition to speeding up visual processing, this approach facilitates the accuracy of visual detections during an ongoing prehensile movement. The common problem with visual detections in cluttered scenes (as the one in Fig. 7 (f)) is that there could be a significant number of false positives after image processing is done. Because we bound visual processing to motor plans of a robot, we significantly reduce false positive detections. In the context of the conducted experiments, there are no false positive detections of the objects in the parts of the visual field that are irrelevant to motor plans, because the relevant objects are not likely to be there.

The computation times presented here are the averages computed from 200 measurements. We have included a supplementary video file which contains the experiments presented here. The video will be available online at http://ieeexplore.ieee.org.

VI. Discussion

In this paper, we have presented one general approach, with two different, but complementary, computational realizations, where visual attention is computed by using modulation signals originating from the robot's motor system. In sharp contrast to the classical approach in computational models of attention and corresponding robotic implementations, where visual attentional relevance is computed based on low-level visual features such as color, edges and intensity contrast, emphasis is put here on tuning the robot vision with respect to the notion of the peripersonal space and forward-planned reaching and grasping movements.

The approach presented here is inspired by the results from psychology and visual neuroscience suggesting that visual attention emerges from the motor system, as elegantly summarized under the premotor theory of attention [58]. The peripersonal space around the body (in both humans and non-human primates) inherently attracts more visual resources

[4]We used the implementation available from the OpenCV library.

that the extrapersonal (beyond reach) space, with and without supporting arm movements [66], [43]. A number of more recent studies with humans show that specialization in the peripersonal space could be additionally fine-tuned in order to support reaching and grasping movements [38].

According to the aforementioned results from the psychology and neuroscience, we have developed two attentional techniques to drive visual processing in humanoid robots: peripersonal space-primed and motor plans-primed models of visual attention. Peripersonal space-primed attention is based on the idea that visual processing supporting reaching and grasping should prioritize the reachable (peripersonal) space of the robot. On the other hand, motor plans-primed attention is constructed around the idea that during movements, the image parts corresponding to the space around motor plans should receive higher priority for visual processing. The peripersonal space-primed attention model is a more general concept and could be used for a variety of applications, including visual exploration of the reachable space, but also during the ongoing movements, as well. Nevertheless, we advocate its use for visual exploration, but not during actual movements, because motor plans-primed attention offers a more specialized framework, which results in higher computational savings. We have taken a machine learning, data-driven exploratory approach to construct the visuomotor transformations and to obtain an implicit notion of the peripersonal space used for guiding visual processing. The benefits of such an approach are that learned models be adapted, if needed, to the visuomotor transformations involving the imperfections of the kinematics and cameras of real robot, and that it overcomes limitations of the classical methods used for representation of the peripersonal space, while still being very efficient to compute (less than a millisecond to compute the outputs of feedforward neural networks). Once the attentional landscape is computed (either peripersonal space-primed or motor plans-primed) it could be used to drive almost any kind of standard image processing technique. We have presented experiments with two popular techniques, with the histogram-based color detector and SURF. For the histogram-based detector, we treat the attentional landscape as the bivariate probability density function and sample locations of the scanning windows by using the Gibbs sampling technique. For SURF, we apply a threshold-based segmentation to constrain computation of SURF features within the parts of the image with higher motor relevance.

Furthermore, in the presented experiments, we have shown how the peripersonal space-primed and motor plans-primed attention can work together. Peripersonal space-primed attention is used to bootstrap initialization of the motor plans-primed attentional mechanism. In order to use motor plans-primed attention, the robot first needs to possess some previous belief where the object might be. This prior information about the object location is used in an iterative procedure: to compute motor plans, which are used to control the robot and for visual updating of the object location by means of motor plans-driven visual processing. The initial guess where the object might be placed could be obtained by first scanning the entire stereo images in the classical way and then proceeding with the iterative procedure until the task ends. However, peripersonal space-primed attention offers a way to constrain the initial visual search, which is a more efficient method than the naive and expensive scanning of the whole images. Once the object to be grasped (and objects to be potentially avoided) is detected, the robot then selects its motor plans, and it switches its visual attentional mechanism to the motor plans-primed, more specialized and more efficient, attentional model that supports visual processing during movements.

Considering the interplay between the motor-primed attentional effects [38] and low-level scene features [18], [21] is a very interesting topic to be assessed in future work, which could bring better biological plausibility to the modeling, as well. We hypothesize that it might be very likely that some weighting scheme between the motor-primed and low-level feature-based attentional mechanisms exists and that the weighting is governed by some high-level, task-aware cognitive inputs. For example, in natural scene exploration tasks, without any motor actions, the weight associated with the low-level feature-based saliency might be greater than the weight corresponding to the motor-primed visual attention. In motor tasks, the importance of these weights is expected to be reversed. Unfortunately, at this point it is not entirely clear how such a hybrid scheme would be beneficial to reaching and grasping. In addition, more insight is needed from biological studies because the investigation of the study of motor-primed visual effects is by itself a very recently established direction in the domain of visual attention. Some early evidence suggests that this integration might exist [67], however, additional work is needed to outline the computational nature of this interaction.

In our modeling, motor-primed attentional landscapes of the left and right camera of the robot are computed independently of each other. The introduced simplification is common to other attentional models, as well (e.g. [68]). In our case, we introduce this simplification for two practical reasons. First, during babbling-like exploratory learning of the workspace to the image planes projection function, due to the fact that both the kinematic chain torso-neck-head and object being visually segmented (i.e. green ball) are moved in an exploratory fashion, the segmented object is significantly more often seen in only one of the cameras in the stereo setup than in both at the same time. Considering the cameras as independent, in our case speeds up data collecting during costly babbling-like exploration for learning. Second, we plan to use monocular pose reconstruction approaches (for example, POSIT) with our humanoid robot. However, to properly address the biological plausibility of motor-primed attention, we would need to take into account the coupling between the eyes in our future work. Such coupling is used in classical feature-based saliency models in the form of introducing the stereo disparity as a feature channel when computing the master saliency map [69]. One interesting direction along the line of biological modeling would be to consider the coupling effects between the binocular rivalry and visual attention mechanisms [70].

Taken together, in this paper, we have shown that our approach can efficiently distribute limited visual resources in a robot system, significantly reducing resources compared to

the classical uniform image processing, but still allowing for a robot to perform complicated tasks, such as manipulation with obstacle avoidance.

REFERENCES

[1] D. Ballard, "Animate vision," *Artificial Intelligence*, vol. 48, no. 1, pp. 57–86, 1991.

[2] M. Wilson, "Six views of embodied cognition," *Psychonomic bulletin & review*, vol. 9, no. 4, pp. 625–636, 2002.

[3] J. S. Werner and L. M. Chalupa, *The visual neurosciences*. Bradford Book, 2004.

[4] J. E. Hoffman and B. Subramaniam, "The role of visual attention in saccadic eye movements," *Perception & Psychophysics*, vol. 57, no. 6, pp. 787–795, 1995.

[5] J. M. Findlay and I. D. Gilchrist, "Eye guidance and visual search," *Eye Guidance in Reading and Scene Perception*, pp. 295–312, 1998.

[6] S. Liversedge and J. Findlay, "Saccadic eye movements and cognition," *Trends in Cognitive Sciences*, vol. 4, no. 1, pp. 6–14, 2000.

[7] M. Hayhoe, A. Shrivastava, R. Mruczek, and J. Pelz, "Visual memory and motor planning in a natural task," *Journal of Vision*, vol. 3, no. 1, 2003.

[8] D. Baldauf, M. Wolf, and H. Deubel, "Deployment of visual attention before sequences of goal-directed hand movements," *Vision Research*, vol. 46, no. 26, pp. 4355–4374, 2006.

[9] D. Baldauf and H. Deubel, "Visual attention during the preparation of bimanual movements," *Vision Research*, vol. 48, no. 4, pp. 549–563, 2008.

[10] W. S. Geisler, "Visual perception and the statistical properties of natural scenes," *Annu. Rev. Psychol.*, vol. 59, pp. 167–192, 2008.

[11] D. Baldauf and H. Deubel, "Attentional selection of multiple goal positions before rapid hand movement sequences: An event-related potential study," *Journal of Cognitive Neuroscience*, vol. 21, no. 1, pp. 18–29, 2009.

[12] B. Noris, J. Keller, and A. Billard, "A wearable gaze tracking system for children in unconstrained environments," *Computer Vision and Image Understanding*, 2010.

[13] M. Land, N. Mennie, J. Rusted *et al.*, "The roles of vision and eye movements in the control of activities of daily living," *Perception*, vol. 28, no. 11, pp. 1311–1328, 1999.

[14] C. Rothkopf, D. Ballard, and M. Hayhoe, "Task and context determine where you look," *Journal of Vision*, vol. 7, no. 14, 2007.

[15] J. Jovancevic-Misic and M. Hayhoe, "Adaptive gaze control in natural environments," *The Journal of Neuroscience*, vol. 29, no. 19, pp. 6234–6238, 2009.

[16] Metta *et al.*, "The icub humanoid robot: An open-systems platform for research in cognitive development," *Neural Networks*, vol. 23, no. 8-9, pp. 1125–1134, 2010.

[17] L. Lukic, A. Billard, and J. Santos-Victor, "Modulating vision with motor plans: A biologically-inspired efficient allocation of visual resources," In Proceedings of the IEEE-RAS International Conference on Humanoid Robots, Humanoids 2013, Atlanta, USA, 2013.

[18] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[19] J. M. Wolfe, "What can 1 million trials tell us about visual search?" *Psychological Science*, vol. 9, no. 1, pp. 33–39, 1998.

[20] P. Reinagel and A. M. Zador, "Natural scene statistics at the centre of gaze," *Network: Computation in Neural Systems*, vol. 10, no. 4, pp. 341–350, 1999.

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

[22] V. Navalpakkam and L. Itti, "Modeling the influence of task on attention," *Vision Research*, vol. 45, no. 2, pp. 205–231, 2005.

[23] S. Frintrop, *VOCUS: A visual attention system for object detection and goal-directed search*. Springer, 2006, vol. 3899.

[24] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artificial intelligence*, vol. 78, no. 1, pp. 507–545, 1995.

[25] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson, "Top-down control of visual attention in object detection," in *Image processing, 2003. icip 2003. proceedings. 2003 international conference on*, vol. 1. IEEE, 2003, pp. I–253.

[26] A. D. Hwang, E. C. Higgins, and M. Pomplun, "A model of top-down attentional control during visual search in complex scenes," *Journal of Vision*, vol. 9, no. 5, p. 25, 2009.

[27] M. Begum and F. Karray, "Visual attention for robotic cognition: a survey," *IEEE Transactions on Autonomous Mental Development*, vol. 3, no. 1, pp. 92–105, 2011.

[28] L. Manfredi, E. S. Maini, P. Dario, C. Laschi, B. Girard, N. Tabareau, and A. Berthoz, "Implementation of a neurophysiological model of saccadic eye movements on an anthropomorphic robotic head," in *IEEE-RAS International Conference on Humanoid Robots*, 2006.

[29] G. Metta, "An attentional system for a humanoid robot exploiting space variant vision," DTIC Document, Tech. Rep., 2001.

[30] F. Orabona, G. Metta, and G. Sandini, "Object-based visual attention: a model for a behaving robot," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 2005.

[31] C. Breazeal, A. Edsinger, P. Fitzpatrick, and B. Scassellati, "Active vision for sociable robots," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 31, no. 5, pp. 443–453, 2001.

[32] L. Aryananda, "Attending to learn and learning to attend for a social robot," in *IEEE-RAS International Conference on Humanoid Robots*. IEEE, 2006, pp. 618–623.

[33] M. W. Doniec, G. Sun, and B. Scassellati, "Active learning of joint attention," in *IEEE-RAS International Conference on Humanoid Robots*. IEEE, 2006, pp. 34–39.

[34] M. Ogino, H. Toichi, Y. Yoshikawa, and M. Asada, "Interaction rule learning with a human partner based on an imitation faculty with a simple visuo-motor mapping," *Robotics and Autonomous Systems*, vol. 54, no. 5, pp. 414–418, 2006.

[35] S. Kestur, M. S. Park, J. Sabarad, D. Dantara, V. Narayanan, Y. Chen, and D. Khosla, "Emulating mammalian vision on reconfigurable hardware," in *IEEE 20th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, 2012.

[36] M. I. Posner, C. R. Snyder, B. J. Davidson *et al.*, "Attention and the detection of signals," *Journal of Experimental Psychology*, vol. 109, no. 2, pp. 160–174, 1980.

[37] C. W. Eriksen and J. D. S. James, "Visual attention within and around the field of focal attention: A zoom lens model," *Perception & Psychophysics*, vol. 40, no. 4, pp. 225–240, 1986.

[38] D. Baldauf and H. Deubel, "Attentional landscapes in reaching and grasping," *Vision Research*, vol. 50, no. 11, pp. 999–1013, 2010.

[39] A. Schiegg, H. Deubel, and W. Schneider, "Attentional selection during preparation of prehension movements," *Visual Cognition*, vol. 10, no. 4, pp. 409–431, 2003.

[40] E. Làdavas, "Functional and dynamic properties of visual peripersonal space," *Trends in cognitive sciences*, vol. 6, no. 1, pp. 17–22, 2002.

[41] A. Maravita and A. Iriki, "Tools for the body (schema)," *Trends in cognitive sciences*, vol. 8, no. 2, pp. 79–86, 2004.

[42] A. J. Rodriguez-Sanchez, E. Simine, and J. K. Tsotsos, "Attention and visual search," *International Journal of Neural Systems*, vol. 17, no. 04, pp. 275–288, 2007.

[43] F. H. Previc, "The neuropsychology of 3-d space." *Psychological bulletin*, vol. 124, no. 2, p. 123, 1998.

[44] F. Maringelli, J. McCarthy, A. Steed, M. Slater, and C. Umilta, "Shifting visuo-spatial attention in a virtual three-dimensional space," *Cognitive Brain Research*, vol. 10, no. 3, pp. 317–322, 2001.

[45] B. J. Losier and R. M. Klein, "Covert orienting within peripersonal and extrapersonal space: Young adults," *Cognitive brain research*, vol. 19, no. 3, pp. 269–274, 2004.

[46] C. L. Reed, J. D. Grubb, and C. Steele, "Hands up: attentional prioritization of space near the hand." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 32, no. 1, p. 166, 2006.

[47] R. A. Abrams, C. C. Davoli, F. Du, W. H. Knapp III, and D. Paull, "Altered vision near the hands," *Cognition*, vol. 107, no. 3, pp. 1035–1047, 2008.

[48] J. D. Cosman and S. P. Vecera, "Attention affects visual perceptual processing near the hand," *Psychological Science*, vol. 21, no. 9, pp. 1254–1258, 2010.

[49] C. C. Davoli, J. R. Brockmole, and A. Goujon, "A bias to detail: how hand position modulates visual learning and visual memory," *Memory & cognition*, vol. 40, no. 3, pp. 352–359, 2012.

[50] L. Fadiga, L. Fogassi, V. Gallese, and G. Rizzolatti, "Visuomotor neurons: ambiguity of the discharge or 'motor' perception?" *International journal of psychophysiology*, vol. 35, no. 2, pp. 165–177, 2000.

[51] P. H. Weiss, J. C. Marshall, G. Wunderlich, L. Tellmann, P. W. Halligan, H.-J. Freund, K. Zilles, and G. R. Fink, "Neural consequences of acting in near versus far space: a physiological basis for clinical dissociations," *Brain*, vol. 123, no. 12, pp. 2531–2541, 2000.

[52] M. F. Rushworth, A. Ellison, and V. Walsh, "Complementary localization and lateralization of orienting and motor attention," *Nature neuroscience*, vol. 4, no. 6, pp. 656–661, 2001.
[53] C. Rothkopf and D. Ballard, "Image statistics at the point of gaze during human navigation," *Visual Neuroscience*, vol. 26, no. 01, pp. 81–92, 2009.
[54] R. Johansson, G. Westling, A. Bäckström, and J. Flanagan, "Eye–hand coordination in object manipulation," *The Journal of Neuroscience*, vol. 21, no. 17, pp. 6917–6932, 2001.
[55] M. F. Land and P. McLeod, "From eye movements to actions: how batsmen hit the ball," *Nature Neuroscience*, vol. 3, no. 12, pp. 1340–1345, 2000.
[56] M. Eimer, J. Van Velzen, E. Gherri, and C. Press, "Manual response preparation and saccade programming are linked to attention shifts: Erp evidence for covert attentional orienting and spatially specific modulations of visual processing," *Brain research*, vol. 1105, no. 1, pp. 7–19, 2006.
[57] H. Deubel and W. X. Schneider, "Attentional selection in sequential manual movements, movements around an obstacle and in grasping," *Attention in action*, pp. 69–91, 2004.
[58] G. Rizzolatti and L. Craighero, "Premotor theory of attention," *Scholarpedia*, vol. 5, no. 1, p. 6311, 2010.
[59] S. Haykin, *Neural Networks: A Comprehensive Foundation (2nd Edition)*. Prentice Hall, 1998.
[60] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly Media, Incorporated, 2008.
[61] S. Kim, A. Shukla, and A. Billard, "Catching objects in flight," *IEEE Transactions on Robotics*, vol. PP, no. 99, pp. 1–17, 2014.
[62] L. Lukic, J. Santos-Victor, and A. Billard, "Learning coupled dynamical systems from human demonstration for robotic eye-arm-hand coordination," IEEE-RAS International Conference on Humanoid Robots (Humanoids), Osaka, Japan, 2012.
[63] L. Lukic, J. Santos-Victor, and A. Billard, "Learning robotic eye–arm–hand coordination from human demonstration: a coupled dynamical systems approach," *Biological cybernetics*, vol. 108, no. 2, pp. 223–248, 2014.
[64] K. P. Murphy, *Machine Learning: a Probabilistic Perspective*, 2012.
[65] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Lecture Notes in Computer Science*, vol. 3951, pp. 404–417, 2006.
[66] M. S. Graziano and C. G. Gross, "The representation of extrapersonal space: a possible role for bimodal, visual-tactile neurons," *The cognitive neurosciences*, pp. 1021–1034, 1995.
[67] H. Deubel, W. X. Schneider, and I. Paprotta, "Selective dorsal and ventral processing: Evidence for a common attentional mechanism in reaching and perception," *Visual Cognition*, vol. 5, no. 1-2, pp. 81–107, 1998.
[68] A. Dankers, N. Barnes, W. F. Bischof, and A. Zelinsky, "Humanoid vision resembles primate archetype," in *Experimental Robotics*. Springer, 2009, pp. 495–504.
[69] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature reviews neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.
[70] J. F. Mitchell, G. R. Stoner, and J. H. Reynolds, "Object-based attention determines dominance in binocular rivalry," *Nature*, vol. 429, no. 6990, pp. 410–413, 2004.

Luka Lukic obtained his B.Sc./M.Sc. in mechatronics and robotics from the University of Novi Sad, Serbia, with First Class Honors, in 2009. He conducted an internship and research for his master thesis at the Vukobratovic Robotics Laboratory, Mihailo Pupin Institute, Belgrade, in 2008-09.

He is currently pursuing a Ph.D. degree in cognitive robotics at the Swiss Federal Institute of Technology in Lausanne, Lausanne (EPFL), Switzerland, and Instituto Superior Técnico (IST), University of Lisbon, Portugal. His research interests are biologically-inspired techniques in robotics, visuomotor coordination in humans and robots, visual attention models, data analysis and machine learning.

Aude Billard is head of the Learning Algorithms and Systems Laboratory (LASA) at the School of Engineering at the EPFL. She received a M.Sc. in Physics from EPFL (1995), a M.Sc. in Knowledge-based Systems (1996) and a Ph.D. in Artificial Intelligence (1998) from the University of Edinburgh.

Her research interests include machine-learning tools to support robot learning through human guidance. This also extends to research on complementary topics, including machine vision and its use in human–robot interaction and computational neuroscience to develop models of motor learning in humans. Aude Billard served as an elected member of the Administrative Committee of the IEEE Robotics and Automation society for two terms (2006-2008 and 2009-2011) and was the recipient of the IEEE-RAS Best Reviewer award. She was a keynote speaker at the IEEE-RAS International Conference on Robotics and Automation (ICRA) in 2013 and at the IEEE International Symposium on Human-Robot Interaction (ROMAN) in 2005, general chair for the IEEE International Conference on Human-Robot Interaction in 2011 and co-general chair for the IEEE International Conference on Humanoid Robots in 2006.

Her research on human-robot interaction and robot programming by demonstration was featured in numerous premier venues (BBC, IEEE Spectrum) and received several best paper awards at major robotics conferences, among which ICRA, IROS and ROMAN.

José Santos-Victor received the Ph.D. degree in Electrical and Computer Engineering in 1995 from Instituto Superior Técnico (IST, Lisbon, Portugal), in the area of Computer Vision and Robotics. He is a Full Professor at the Department of Electrical and Computer Engineering of IST and a researcher of the Institute of Systems and Robotics (ISR), at the Computer and Robot Vision Laboratory (VISLAB).

He is the scientific responsible for the participation of IST in various European and National research projects in the areas of computer vision and robotics. His research interests are in the areas of computer and robot vision, particularly in the relationship between visual perception and the control of action, biologically inspired vision and robotics, cognitive vision and visual controlled (land, air and underwater) mobile robots.

Prof. Santos-Victor is an IEEE member and was an associate editor of the IEEE Transactions on Robotics.