Autonomous Switching of Top-down and Bottom-up Attention Selection for Vision Guided Mobile Robots

Tingting Xu[1], Nikolay Chenkov[1], Kolja Kühnlenz[1,2] and Martin Buss[1]

[1]Institute of Automatic Control Engineering (LSR)
[2]Institute for Advanced Study (IAS)
Technische Universität München
D-80290 Munich, Germany
Email: {tingting.xu, kolja.kuehnlenz, m.buss}@ieee.org, nikolay.chenkov@mytum.de

Abstract— In this paper an autonomous switching between two basic attention selection mechanisms, top-down and bottom-up, is proposed, substituting manual switching. This approach fills the gab in object search using conventional top-down biased bottom-up attention selection: the latter one fails, if a group of objects is searched whose appearances can not be uniquely described by low-level features used in bottom-up computation models. Two internal robot states, observing and operating, are included to determine the visual selection behavior. A vision guided mobile robot, equipped with an active stereo camera, is used to demonstrate our strategy and evaluate the performance experimentally.

I. INTRODUCTION

To achieve an efficient processing of visual information about the environment, humans select their focus of visual attention, such that the most interesting regions will be processed first in detail. Studies about human visual perception show that visual attention selection is affected by two distinct types of attentional mechanisms: top-down and bottom-up. Top-down signals are derived from the task specification or previous knowledge and highlight the task-relevant information. Top-down attention selection is straightforward and efficient for task accomplishment. In contrast, bottom-up attention selection is inspired by a neuronal architecture of early primate vision. It is induced by stimuli regarding color, intensity, orientation etc. on several hierarchy levels. Without concrete top-down information, pure bottom-up is the only way to select potentially important information of the environment for further processing. In human attention systems, top-down and bottom-up selection always work together to determine the attentional allocation and control the human gaze behavior.

Operating in the real world, a robot has normally a task such as detecting and manipulating a target object. For a mobile robot, a typical task is to find a target and move toward it. In a simple scenario with unique target objects, a conventional top-down biased bottom-up strategy can help a lot in terms of efficiency [1]. However, it fails, if a group of objects is searched whose appearances can not be uniquely described by low-level features used in a bottom-up computation model. For example, several different traffic signs are all salient in color but in different geometry and with different text on them. They are, therefore, not distinguishable from

each other only relying on low-level features used in bottom-up attention selection. In this instance, top-down information is ineffectual and a bottom-up attention selection is necessary to initialize the object search process. However, during task performance, task-oriented attention selection is essential for efficiency. Especially if there is no top-down relevant object in the field of view, a pure top-down attention selection can also use position data in 3D task-space, while bottom-up or top-down biased bottom-up attention selection only relies on 2D image data. Therefore, a switching between top-down based state and bottom-up based state is proposed to deal with different situations, which enables autonomy of robots in terms of visual behavior. This paper is the first attempt of an autonomous switching between these two kinds of attention selection strategies and fills the gap for object search not solvable using conventional combination of them. A vision-guided mobile robot (see Fig. 6), the Autonomous City Explorer (ACE) [2] developed at our institute, is used to demonstrate our strategy and evaluate the performance experimentally. It is equipped with an active vision system, consisting of a Bumblebee XB3 stereo camera from Point Grey Research Inc. and a high-performance pan-tilt platform [3].

The paper is organized as follows: In Section II, related works about combination of top-down and bottom-up attention selection are introduced. In Section III, the proposed autonomous switching strategy is presented. In Section IV, the performance of our strategy is experimentally demonstrated. The results are shown and discussed. Conclusions are given in Section V.

II. RELATED WORK

In the last few decades, bottom-up saliency-based attention selection models have also become focus of robot view direction and attention planning. A computational model, the saliency map model, was firstly proposed in [4] and further developed by [5] and [6]. In this model the salient positions in a static image are selected by low-level features. The saliency map predicts a human-like visual attention allocation. A saliency-driven vision system has already been applied on a robot head [7], which uses a bottom-up visual

attention mechanism to focus on interesting objects in the environment.

To achieve an efficient task accomplishment, task-relevant top-down factors can be integrated into bottom-up attention selection models to bias the visual attention selection. To solve the problem of visual search for a given target in an arbitrary 3D space for robot vision systems, the probability of finding the target is optimized in [8], given a fixed cost limit in terms of total number of robotic actions the robot needs to find its visual target, facilitated by attentive processes. A complex object recognition system on a mobile robot is proposed in [9], which is capable of locating numerous challenging objects amongst distractors. The potential objects are ranked using a bag-of-features technique and identified using an attention mechanism in a limited time. In [10] an approach to an optimal gaze control system for autonomous vehicles is proposed in which the perceptive situation and subjective situation are also predicted. In [1] an environment adapted active multi-focal vision system is proposed. A top-down biased bottom-up attention selection strategy without previous training is applied. A Kalman-filter is used to estimate the weights of feature maps in building a task-relevant saliency map. The saliency of top-down elements and the saliency of bottom-up components are combined in [11] in a way that the top-down part is initialized by the bottom-up part, hence resulting in a selection of the behaviors to deal with the limited computational resources. In [12] a biologically motivated computational attention system VOCUS is introduced, which has two operation modes: the exploration mode based on strong contrasts and uniqueness of a feature and the search mode using previously learned information of a target object to bias the saliency computations with respect to the target. In [13] a salient proto-object detection model based on selective visual attention is suggested in the way that the objects are attended to before recognized. In [14] a task-driven object-based visual attention model for robot applications is proposed, which involves five components: pre-attentive object based segmentation, bottom-up still attention, bottom-up motion attention, top-down object-based biasing and contour based object representation. Task-specific moving object detection and still object detection are operated based on this model. Up to now, switching of top-down and bottom-up attention has only been proposed in [12] and [15]. However, the switching in those systems is activated by users manually.

To perform the switching autonomously, we define transition conditions to trigger the switching from top-down to bottom-up and from bottom-up to top-down, in order to realize an autonomous visual attention selection.

III. AUTONOMOUS SWITCHING BETWEEN TOP-DOWN AND BOTTOM-UP ATTENTION SELECTION

Fig. 1 illustrates the switching mechanism of attention selection. The robot has two modes, namely observing mode and operating mode. Two attention selection states, a top-down state and a bottom-up state, as well as 4 transitions (*bt,*

tb, tt, bb) between the states are contained in the observing mode. Because a robot is normally assigned with a specific manipulation or navigation task, not just looking around for the target object, an operating mode is considered beside the observing mode. In operating mode, the robot accomplishes its task, such as approaching or manipulating an object, using top-down task-oriented attention selection. In this section we discuss where the robot should attend to in each state and how the autonomous switching between the states is conducted.
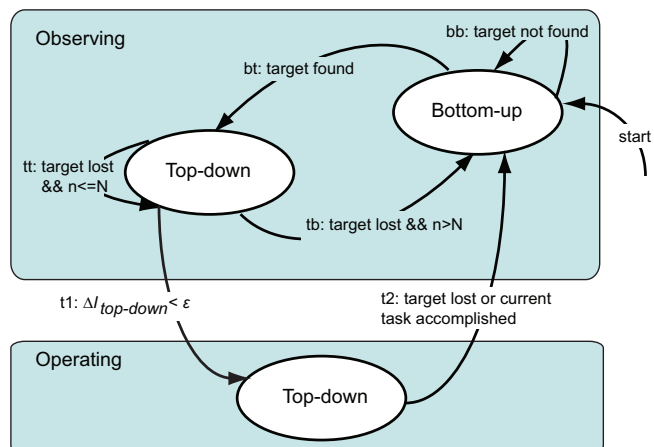

Fig. 1.   Finite state machine of the switching mechanism.

A. Bottom-up State

In bottom-up state, the robot focuses on a salient area in the field of view. Since the goal of this paper is to solve the problem that target objects are not uniquely described or that top-down information with low-level features used in bottom-up attention selection model is not available, a bottom-up based attention selection model is used to select candidate regions which may contain target objects. We use a well-known standard computational model for bottom-up attention selection, namely the saliency map model proposed in [4].

In Fig. 2 the saliency map model is visualized. An input image of e.g. 640×480 pixels is sub-sampled into dyadic Gaussian pyramids in three channels (intensity, orientation for $0°, 45°, 90°, 135°$, opponent color in red/green and blue/yellow). The size of the image is reduced from 640×480 to 320×240, ..., and to 2×1 successively in each lower level. Then center-surround differences are calculated for the images in the Gaussian pyramids. In this phase feature maps are generated in which distinctive pixels with respect to their neighborhood are highlighted. Using across-scale combinations the feature maps are combined and normalized into a conspicuity map in each channel. A saliency map is a linear combination of the conspicuity maps. The bright pixels in the saliency map are the salient and interesting pixels predicted by this model.

An information based extension of this model is made to describe the influence of temporal novelty on the total
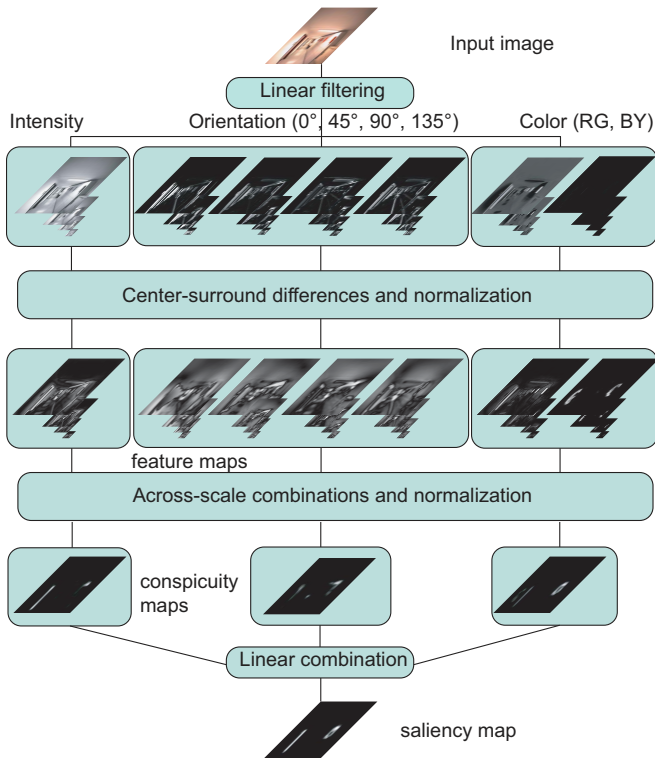
Fig. 2. Saliency map model

preference of the regions selected by the saliency map. For the temporal novelty we apply a Bayesian definition of the information content of an image directly using the saliency map. The notion "surprise" is used here to indicate the unexpected events [16]. Only the positions being spatially salient and temporally surprising are taken to draw the robot's attention. We build a surprise map using two consecutive saliency maps without camera movement to find the unexpected events.

To achieve this, we model the data D received from the saliency map as Poisson distribution $M(\lambda(x_i, y_i))$, where $\lambda(x_i, y_i)$ stands for the saliency value of the pixel (x_i, y_i). Therefore, a prior probability distribution $p_i(x_i, y_i)$ can be defined as a Gamma probability density for the i-th pixel:

$$p_i(x_i, y_i) = \gamma(\lambda, \alpha, \beta) = \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)}, \quad (1)$$

with the shape $\alpha > 0$, the inverse scale $\beta > 0$, and $\Gamma(\cdot)$ the Euler Gamma function.

The posterior probability distribution $p((x_i, y_i)|D)$ is obtained from the second saliency map with the new saliency value $\lambda'(x_i, y_i)$. The parameters α and β are supposed to change into α' and β', according to

$$\alpha' = \xi\alpha + \lambda', \quad \text{and} \quad \beta' = \xi\beta + 1, \quad (2)$$

with a forgetting factor ξ, $0 < \xi < 1$.

Then, a surprise map with surprise value τ is estimated as the Kullback-Leibler divergence between the prior probability distribution $p_i(x_i, y_i)$ and the posterior probability

distribution $p_i(x_i, y_i|D)$ as follows:

$$\tau(x_i, y_i) = KL(p_i(x_i, y_i)||p_i(x_i, y_i|D)). \quad (3)$$

Finally, the pixel coordinate (x_*, y_*) with the maximum surprise value is found for the robot gaze control

$$(x_*, y_*) = \arg\max_{(x_i, y_i)} (\tau(x_i, y_i)). \quad (4)$$

In the example shown in Fig. 3, the rectangles in solid lines are the attention focus predicted by the surprise map. In the left column, a moving human is selected as the focus of attention because of its high surprise value. In bottom-up state, the robot attends to the image region limited by the rectangle in solid lines, although no robot task such as human detection is assigned to the robot. The focus of attention (the masked image region) and the most salient/surprising position (the rectangle) indicate the same position. More examples of the surprise map can be found in [17]. In bottom-up state the salient/surprising image regions in the input image are viewed sequentially according to their saliency/surprise value.

B. Top-down State

In top-down state, robot concentrates itself on image region containing task-relevant information. The conventional robot tasks can be approaching, avoiding or grasping an object in which the position estimation of the object is the main objective. To perform this task, the robot should attend to the region which contains the target object to get a better accuracy.

In Fig. 3 the right column shows an example for top-down state. A robot is supposed to detect a traffic sign and approach it. The region around a target object, the masked region in the right-bottom image, is selected as the current robot attention focus and is further processed in detail, although this region is not the most salient/surprising region at this moment, namely the region in the rectangle. In short, in top-down state, the position of the target object is known. No matter how salient and surprising the other features are, to perform its task, the robot attends to the detected target object.

C. Switching Mechanism

The main contribution of this paper is to realize an autonomous switching between top-down based and bottom-up based visual attention selection considering robot task performance. The transition conditions are defined as follows.

After initialization, the image region to be further processed is selected in bottom-up state of observing state, since the position of the target object is unknown at this moment. Once a target is found in the selected region, top-down state is activated. The image region around the target is selected constantly, ignoring the other salient features. If the target is lost, for example due to lighting condition change or humans and vehicles hiding the target object, the robot should continue focusing on the last region for N frames at first to see if the target object is re-detectable. If the robot

Fig. 3. Left column: attention selection in bottom-up state; Right column: attention selection in top-down state; Top row: original input images; Bottom row: the resultant images; Rectangle in solid lines: the salient/surprising image region; Masked region: the current focus of attention; Circle: detected target object.

stays in top-down state for n frames, $n \geq N$, and the target is still unseen, bottom-up selection is triggered to search for the previous target.

If the observation of the target object in top-down state is accurate enough, the robot starts to operate. To evaluate the accuracy of the observation, we model the m-dimensional system state $\underline{x} \in \mathbb{R}^m$ of the current robot task as a 2D Gaussian distribution with mean value $\underline{\mu}$ and covariance matrix $R_{\underline{x}}$ in the task-space computed using a Kalman-filter. The system state $\underline{x}$ is chosen according to the current task and can be robot position and velocity for a self-localization task or object position and velocity for an object tracking task. The distribution at the previous time step is regarded as the prior probability density function (pdf) $p(\underline{x})$, while the pdf at the current time step is $q(\underline{x})$ with a continuous variable $\underline{x}$ for specific tasks. Both of them are defined as follows:

$$
\begin{aligned}
p(\underline{x}) &= \frac{1}{(\sqrt{2\pi})^m (\det R_{\underline{x}}^{k-1})^{1/2}} \cdot \\
&\quad \exp(-\frac{1}{2}(\underline{x}^{k-1} - \underline{\mu}^{k-1}) \cdot \\
&\quad (R_{\underline{x}}^{k-1})^{-1} \cdot (\underline{x}^{k-1} - \underline{\mu}^{k-1})^T), \quad (5)
\end{aligned}
$$

and

$$
\begin{aligned}
q(\underline{x}) &= \frac{1}{(\sqrt{2\pi})^m (\det R_{\underline{x}}^{k})^{1/2}} \cdot \\
&\quad \exp(-\frac{1}{2}(\underline{x}^{k} - \underline{\mu}^{k}) \cdot \\
&\quad (R_{\underline{x}}^{k})^{-1}(\underline{x}^{k} - \underline{\mu}^{k})^T), \quad (6)
\end{aligned}
$$

with the dimension m of the state variable $\underline{x}$ and the time step k.

The relative entropy is then computed as follows:

$$
\begin{aligned}
\Delta I_{top-down} &= KL(p(\underline{x})||q(\underline{x})) \\
&= \int_{-\infty}^{\infty} p(\underline{x}) \log \frac{p(\underline{x})}{q(\underline{x})} d\underline{x} \quad \text{in [bit]}. \quad (7)
\end{aligned}
$$

We define an empirical threshold ε for the relative entropy $\Delta I_{top-down}$ between the predicted and the updated state estimate as one of the criteria for evaluating the observation accuracy. The less $\Delta I_{top-down}$ is, the less the estimation and its expected value vary, and therefore, the better is the position estimation. If the information measure at the k-th step is smaller than this threshold, the observation at this step is regarded as successfully executed. Upon this value the robot takes the decision what action to perform next: operating or observing. Respectively, if the task is finished or the target is lost, the robot stops the current operation, turns into bottom-up state and observes.

IV. PERFORMANCE EVALUATION

To demonstrate our strategy, experiments were conducted using the ACE robot mentioned in Section I.

A. Experiment

Fig. 4 shows the experimental scenario in our laboratory. The ACE robot was supposed to detect three different signs one after another. The positions of the signs were unknown. Once a sign was detected and the position of this sign was satisfyingly estimated, ACE moved straight ahead and tracked the sign using the active camera head during the movement, until it reached the position one meter in front of the sign. Then, the head of the robot should turn to another direction randomly and search for another sign and so on.

Fig. 4. Experiment setup.

These three signs can not be uniquely described by low-level features used in the saliency map model and therefore can not be easily recognized and distinguished using a top-down biased bottom-up attention selection. For object recognition we use previously trained classifiers based on Haar-like features [18]. To lower the computational cost of object recognition, the classifiers were only applied in the focus of attention selected in top-down state and bottom-up

state. The whole input image represents a peripheral sensor input, while the focus region represents a foveated sensor input with higher resolution.

B. Results

Fig. 6 illustrates the experimental results. Images with attention focus region (the masked region) and salient/surprising region (the region in the rectangle in solid lines) as well as the frame number are shown. At the first step, ACE looked straight ahead and bottom-up state was activated. In frame 1, the blue sign was detected. The focus of attention changed into top-down state. The image region around the blue sign was selected in the following frames, until the robot reached the position one meter in front of the blue sign (frame 44). The threshold for $\Delta I_{top-down}$ was set to be 0.12 bit. Then, the robot turned its head randomly to the right side and detected the yellow sign coincidentally (frame 45). The robot still stayed in top-down state. After the position estimation was satisfyingly accomplished, the robot started to move and track the yellow sign. In frame 111 the sign was lost and bottom-up state was activated after several frames. In frame 127, the yellow sign was re-detected in the image region selected in bottom-up state. Top-down state was triggered again. After the robot reached the position one meter in front of the yellow sign, the head was randomly directed and the state was bottom-up state again (frame 149). In frame 151 and 214 the red sign was detected and tracked. For 228 frames in total, there are 18 frames in bottom-up state and 210 frames in top-down state. Fig. 5 illustrates the evolution of the relative entropy $\Delta I_{top-down}$ and the switching between top-down and bottom-up state. The semitransparent time intervals indicate the operating state, in which the robot was moving.
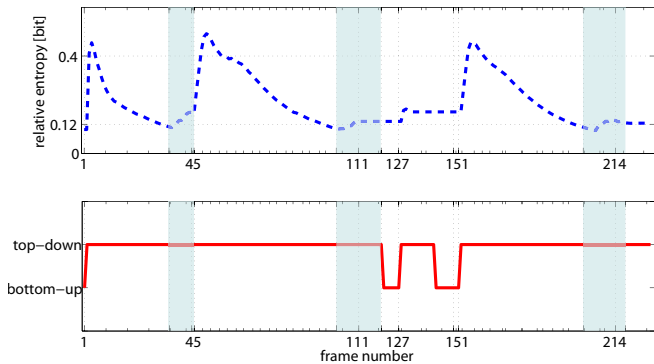

Fig. 5. Relative entropy evolution and the respective attention control scheme. The semitransparent time intervals indicate the operating state.

The experiment is also shown in the short accompanying video. To evaluate the visual guidance performance separately, the other sensors on ACE such as laser range finders were deactivated. To avoid possible crashes with the signs, we set a very low value to $\Delta I_{top-down}$, which caused a relatively long observing period before the robot started to operate. However, this can be easily improved if other sensor modalities are used for obstacle avoidance as well.

Tab. I shows the average computation time which was taken in different phases. Since the bottom-up attention selection was implemented on Graphics Processing Units (GPUs) [19], real-time processing in this part is ensured. The expensive processing is due to the object recognition algorithm. There is a large improvement in the performance if the robot searches for the signs only in the attention focus but not in the whole image.

Task	Time
Image capture	67 ms
Surprise map computation	20 ms
+ Search for a sign in attention focus	31 ms
+ Search for 3 signs in attention focus	33 ms
- Search for a sign in the whole image	183 ms
- Search for 3 signs in the whole image	373 ms

TABLE I
Average computation time for each step in the experiment.

C. Discussion

In this experiment, the searched targets, namely three different signs, have different appearances. However, it is impossible to use uniform or similar model parameters such as the weights of feature maps in bottom-up attention selection models to represent and distinguish them. Pure bottom-up attention facilitates the robot task accomplishment by providing attention focus candidates and reducing the detection time.

In our experiment, the resolution of the vision sensor is still sufficient for the sign recognition. If more resolution is required to further process the selected region, bottom-up state provides potential image region candidates before a target object is detected and is a must for an efficient utilization of high-resolution cameras [1]. Otherwise, the high-resolution camera has to search objects in the environment randomly and inefficiently.

To accelerate the whole task performance, it is obvious that the pure bottom-up attention selection should be active as less as possible, although the bottom-up state is necessary. Three solutions are suggested:

- We can reduce the computation time for bottom-up state, which has already been achieved using multi-GPU implementation [19].
- If an object is found, features related to bottom-up attention selection should be saved. If the object is just lost, a top-down biased bottom-up mode [1] can be used for a more efficient search.
- Inhibition of return is applied to avoid repeated view of the positions which have already been observed.

V. CONCLUSIONS

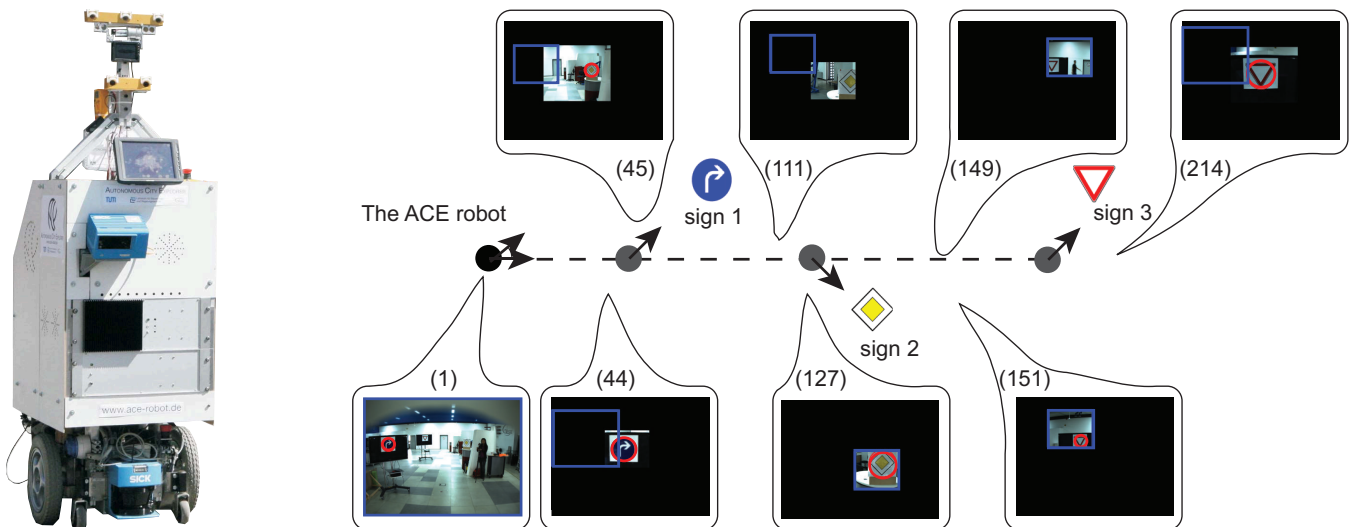In this paper a switching between top-down state and bottom-up state is proposed to deal with scenarios in which

Fig. 6. Left: The Autonomous City Explorer (ACE) robot comprising an active vision system and a passive stereos camera (not used in this paper). Right: The experimental results comprising images of robot attention focus and frame number. The solid circles: the ACE robot. The arrows on the robot: the view direction of the active camera head. The dashed line: the robot trajectory.

a group of target objects are searched which cannot be uniquely represented by low-level features used in bottom-up attention selection model. This is the first attempt of an autonomous switching between top-down and bottom-up attention selections and fills the gap for object search with the problems mentioned above. A vision-guided mobile robot ACE, equipped with an active vision system, is used to demonstrate our strategy and evaluate the performance experimentally. The necessity and efficiency of this autonomous switching are demonstrated.

The strategy seems intuitive and straightforward. However, this capability of autonomous switching of visual attention selection models enables a vision-guided mobile robot to be "autonomous" in terms of visual behavior. The selection of attention focus is adapted to the internal robot state, observing or operating.

REFERENCES

[1] T. Xu, H. Wu, T. Zhang, K. Kühnlenz, and M. buss. *Environment Adapted Active Multi-focal Vision System Using Kalman Filter for Object Detection*. In Proc. Int. Conf. Robotics and Automation, 2009.

[2] A. Bauer, K. Klasing, G. Lidoris, Q. Mühlbauer, F. Rohrmüller, S. Sosnowski, T. Xu, K. Kühnlenz, D. Wollherr and M. Buss. *The Autonomous City Explorer: Towards Natural Human-Robot Interaction in Urban Environments*. International Journal of Social Robotics 1, no. 2, 127-140, 2009.

[3] K. Kühnlenz, M. Bachmayer and M. Buss, *A Multi-Focal High-Performance Vision System*. In Proc. Int. Conf. Robotics and Automation, 2006.

[4] L. Itti, C. Koch and E. Niebur. *A model of saliency-based visual attention for rapid scene analysis*. In Pattern Analysis and Machine Intelligence, vol. 20, 1254-1259, 1998.

[5] H.Yee, S. Pattanaik and D. P. Greenberg. *Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments*. In ACM Transactions on Graphics. ACM Press, 39-65, 2001.

[6] W. J. Won, S. W. Ban and M. Lee. *Real Time Implementation of a Selective Attention model for the Intelligent Robot with Autonomous Mental Development*. In Proc. Int. Symp. Industrial Electronics, 2005.

[7] S. Schaal and L. Itti. *Learning and Attention with a Humanoid Robot Head*. USC, Los Angeles, USA, 2005.

[8] J. K. Tsotsos, K. Shubina. *Attention and Visual Search: Active Robotic Vision Systems that Search*. In Proc. the 5th Int. Conf. Computer Vision Systems, 2007.

[9] P. E. Forssen, D. Meger, K. Lai, S. Helmer, J. J. Little, D. G. Lowe. *Informed Visual Search: Combining Attention and Object Recognition*. In Proc. Int. Conf. Robotics and Automation, 2008.

[10] M. Pellkofer and E.D. Dickmanns. *EMS-Vision: Gaze Control in Autonomous Vehicles*. In Proc. IEEE Intelligent Vehicles Symposium 2000.

[11] B. Khadhouri and Y. Demiris, *Compound effects of Top-down and Bottom-up influences on Visual Attention during Action Recognition*. In Proc. the 19th Int. J. Conf. Artificial Intelligence, pp. 1458-1463, 2005.

[12] S. Frintrop. *VOCUS: A Visual Attention System for Object Detection and Goal-directed search*. PhD thesis, University of Bonn, 2006.

[13] D. Walther and C. Koch. *Modeling attention to salient proto-objects*. ScienceDirect. Neural Networks 19, 1395-1407, 2006.

[14] Y. Yu, G. K. I. Mann and R. G. Gosine. *A task-driven object-based attention model for robots*. In Proc. Int. Conf. Robotics and Biomimetics, 2007.

[15] C. Breazeal and B. Scassellati (1999). *A context-dependent attention system for a social robot*. In Proc. of the 16th Int. J. Conf. Artificial Intelligence, 1146–1151, 1999.

[16] L. Itti and P. F. Baldi. *Bayesian Surprise Attracts Human Attention*. Advances in Neural Information Processing Systems, Vol. 19 (NIPS*2005), pp. 547-554, 2006.

[17] T. Xu, Q. Mühlbauer, S. Sosnowski, K. Kühenlenz and M. Buss. *Looking at the Surprise: Bottom-Up Attentional Control of An Active Camera System*. In Proc. the 10th Int. Conf. Control, Automation, Robotics and Vision, 2008.

[18] Q. Mühlbauer, S. Sosnowski, T. Xu, T. Zhang, K. Kühnlenz, and M. Buss. *The Autonomous City Explorer Project: Towards Navigation by Interaction and Visual Perception*. In Proc. Int. Conf. Robotics and Automation, 2009.

[19] T. Xu, T. Pototschnig, K. Kühnlenz and M. Buss. *A High-Speed Multi-GPU Implementation of Bottom-Up Attention Using CUDA*. In Proc. Int. Conf. Robotics and Automation, 2009.