[38] J. S.-C. Yuan, "A general photogrammetric method for determining object position and orientation," *IEEE Trans. Robot. Autom.*, vol. 5, no. 2, pp. 129–142, Apr. 1989.

[39] É. Marchand, F. Spindler, and F. Chaumette, "VISP for visual servoing: A generic software platform with a wide class of robot control skills," *IEEE Robot. Autom. Mag.*, vol. 12, no. 4, pp. 40–52, Dec. 2005.

Autonomous Behavior-Based Switched Top-Down and Bottom-Up Visual Attention for Mobile Robots

Tingting Xu*, Student Member, IEEE*, Kolja Kühnlenz*, Member, IEEE*, and Martin Buss*, Member, IEEE*

***Abstract*—In this paper, autonomous switching between two basic attention selection mechanisms, i.e., top-down and bottom-up, is proposed. This approach fills a gap in object search using conventional top-down biased bottom-up attention selection, which fails, if a group of objects is searched whose appearances cannot be uniquely described by low-level features used in bottom-up computational models. Three internal robot states, such as observing, operating, and exploring, 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. This approach facilitates adaptations of visual behavior to different internal robot states and benefits further development toward cognitive visual perception in the robotics domain.**

***Index Terms*—Vision-guided robotics, visual attention control.**

I. Introduction

To achieve efficient processing of visual information about the environment, humans select their focus of attention (FOA), 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 mechanisms: top-down and bottom-up. Top-down signals are derived from the task specification or the previous knowledge and highlight the task-relevant information. It is goal-directed and essential for task accomplishment. In contrast, bottom-up attention is driven by distinct stimuli based on primary visual features. Interaction and coordination of both enable gaze-fixation-point selection and guide the visual behavior. To deal with the limited processing capability of most technical systems, especially autonomous mobile robots, a biologically plausible and technically applicable visual attention system is to be developed in order to fill the gap between the fundamental studies and the robotics research.

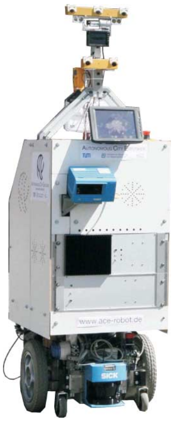

Fig. 1. ACE robot.

Normally, when operating in the real world, a robot has 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 cannot be uniquely described by low-level features used in a primary bottom-up computation model. For example, different traffic signs are all salient in color but different in geometry and have different patterns on them. They are, therefore, not distinguishable from each other and only rely on low-level features used in bottom-up attention selection. An exhaustive search is still needed. To lower the computational cost, a search window is usually defined for exhaustive search as the robot FOA, in which the exhaustive search is conducted.

A search window based on bottom-up attention can predict image regions with higher probability to contain a target object, while a search window based on top-down attention is efficient for task accomplishment. Both bottom-up attention and top-down attention are essential for robot-attention control. On the one hand, if a task-relevant object is not located in the robot field of view (FOV), pure top-down attention selection can also use position data in the 3-D task space to direct robot attention toward the target, while bottom-up or top-down biased bottom-up attention selection only relies on the 2-D image data. On the other hand, if there is no task-relevant information in the FOV at all, pure bottom-up attention can guide the robot attention to explore the environment in a flexible way. In this paper, autonomous switching between top-down and bottom-up attention mechanisms is proposed, which enables autonomy of robots in terms of adaptations of visual behavior to different internal robot states and which fills the gap for object search not solvable using conventional combination of them. A vision-guided mobile robot, which is the Autonomous City Explorer (ACE) [2] developed at our institute (see Fig. 1), is used to demonstrate our strategy and evaluate the performance experimentally. It is equipped with an activevision system, which consists of a Bumblebee XB3 stereo camera from Point Grey Research, Inc., and a high-performance pan-tilt platform [3].

This paper is organized as follows: In Section II, related works about combination of top-down and bottom-up attention selections are introduced. In Section III, the proposed autonomous switching

Manuscript received January 25, 2010; revised May 31, 2010; accepted July 25, 2010. Date of publication August 26, 2010; date of current version September 27, 2010. This paper was recommended for publication by Associate Editor T. Kanda and Editor G. Oriolo upon evaluation of the reviewers' comments. This work was supported in part by the German Research Foundation (DFG) Excellence Initiative Research Cluster *Cognition for Technical Systems* (CoTeSys) (www.cotesys.org) and in part by the Institute for Advanced Study, Technische Universität München (www.tum-ias.de).

T. Xu and M. Buss are with the Insititute of Automatic Control Engineering, Technische Universität München, Munich 80290, Germany (e-mail: tingting.xu@ieee.org; m.buss@ieee.org).

K. Kühnlenz is with the Insitiue of Automatic Control Engineering, Technische Universität München, Munich 80290, Germany, and also with the Institute for Advanced Study, Technische Universität München, Munich 80333 Germany (e-mail: k.kuehnlenz@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TRO.2010.2062571

strategy is presented. In Section IV, the performance of our strategy is experimentally demonstrated, and results are shown and discussed. Conclusions and future work are given in Section V.

II. Related Work

Over the past few decades, bottom-up saliency-based attention-selection models have also become focus of robot view direction and attention planning. Based on fundamental findings in cognitive psychology and neuroscience [4]–[6], computational models for visual attention selection have been proposed [7]–[13]. In the computer-vision domain, visual attention can facilitate object detection [14]–[17], segmentation [18], tracking [19]–[21], and intention understanding [22], [23].

In the robotics domain, visual attention is used to demonstrate current visual interest of robots on the one hand. Many active visual attention systems have been proposed [24], [25]. One of the earliest implementations of visual attention on mobile robots is introduced in [26]. A camera is mounted passively on a mobile robot. A segregation of visual stimuli based on *connectionist model* by means of synchronization of spiking neurons is used to bind image features that correspond to objects. Then, the largest one of the segregated objects is selected and approached by the robot. Although only edge features are used, this system exhibits a primary version of visual attention of mobile robots. In [27], a saliency-driven vision system is also applied to a robot head, which uses a bottom-up visual-attention mechanism to focus on interesting objects in the environment in real time and attends to them. Visual attention for eye and head animation of a realistic virtual human head is applied in [28], using an extended version of the neurobiological model proposed in [10]. In this animation, flicker features and motion features are included as well to deal with the temporal changes and moving objects. Moreover, coordination of eye and head movement is also concerned to achieve a realistic animation and rendering. Various features such as color, intensity, edges, stereo, and motion are used in [29] to drive the gaze of a humanoid head toward potential regions of interest. In [30], a stereo saliency map is used for a vergence control system. In [31], another multifocal-camera system is presented, which consists of foveal vision and peripheral vision. This system is able to locate and recognize objects in the real world, using the top-down object characteristics: hue saliency and 3-D size. The attentional process is performed in a relatively wide FOV, while recognition is conducted in the high-resolution foveal center. Active gaze control for visual simultaneous localization and mapping using features detected by an attention system is applied in [32], which supports the system with tracking, redetection, and exploration behaviors.

To achieve an efficient task accomplishment, task-relevant top-down factors can be integrated into bottom-up attention models to bias the visual attention selection. In [33], the weights of top-down and bottom-up factors are combined, in which an offline optimization of the top-down weights and a context learning based on neural network are conducted using a large set of examples. The balance between top-down and bottom-up is not fixed. The integration of a simple context vector can achieve improvement for object-searching task. To solve the problem of visual search for a given target in an arbitrary 3-D space for robot-vision systems, the probability of finding the target is optimized in [34], given a fixed cost limit in terms of total number of robotic actions the robot needs to find its visual target, which is facilitated by attentive processes. A complex object-recognition system on a mobile robot is proposed in [35], which is capable of locating numerous challenging objects among distractors. The potential objects are ranked using a bag-of-features technique and identified using an attention mechanism in a limited time.

Only a few works have up to now considered switching between top-down and bottom-up visual behavior. In [36], top-down object search and bottom-up environment exploration using the same saliency-map model and robot platform are proposed. However, the switching between them is manual. In [23], a top-down part is initialized by a bottom-up part to recognize actions, track the actions, and determine the current context. In [37], visual attention is switched between different targets. Instead of a pure bottom-up or top-down state, visual attention allocation is determined by a reasonable weighting between top-down and bottom-up signals to demonstrate the robot gaze preference. In [38], a task-driven object-based visual-attention model for robot applications is proposed that involves five components: preattentive 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. In [35], three visual behaviors are defined: Exploration behavior, coverage behavior, and viewpoint-selection behavior. The first behavior is more a robot exploration behavior than a visual behavior. In the second behavior, potential objects are explored by the peripheral vision using bottom-up attention. After the environment is fully covered, novel perspectives of the objects are captured and the object recognition is conducted in viewpoint-selection behavior. The top-down state has been started only once. Strictly speaking, none of the aforementioned works have applied autonomous switching between pure top-down and bottom-up attention mechanisms.

Moreover, most visual-attention systems are studied decoupledly, where a goal-directed robot operation is commonly ignored. A robot should always be supposed to do something with the target object, such as approaching or manipulating. Considering this, robot visual-attention behavior should be adapted to the internal robot state to achieve a complete system.

Therefore, switching between top-down visual state and bottom-up visual state is proposed here, which enables autonomy of robots in terms of visual behavior. This autonomous switching between these two kinds of attention-selection mechanisms is also adapted to different internal robot states and fills the gap for object searches that are not solvable using a conventional combination of them.

III. Autonomous Switching of Attention Mechanisms

The switching mechanism of attention selection for an autonomous robot is illustrated in Fig. 2, which is an extension from our previous work [39]. Three different robot internal modes are considered:

1) *Exploring mode*: The robot has no specific task and just explores the world by looking at interesting parts of the environment;
2) *Searching mode*: The robot has a specific task and searches for its current target object;
3) *Operating mode*: The robot is accomplishing its task, e.g., moving to or manipulating the detected target object.

Four attention-selection states are assigned to the robot visual behavior: the bottom-up state in the exploring mode (abbreviated as BU_e), the bottom-up state in the searching mode (abbreviated as BU_s), the top-down state in the searching mode (abbreviated as TD_s), and the top-down state in the operating mode (abbreviated as TD_o). Seven transitions are defined. In this section, it is discussed how the robot FOA is determined in each state and how the autonomous switching between the states is conducted.

A. Bottom-Up State

In the bottom-up state, the robot focuses on an interesting area in the FOV. A bottom-up-based attention-selection model is used to

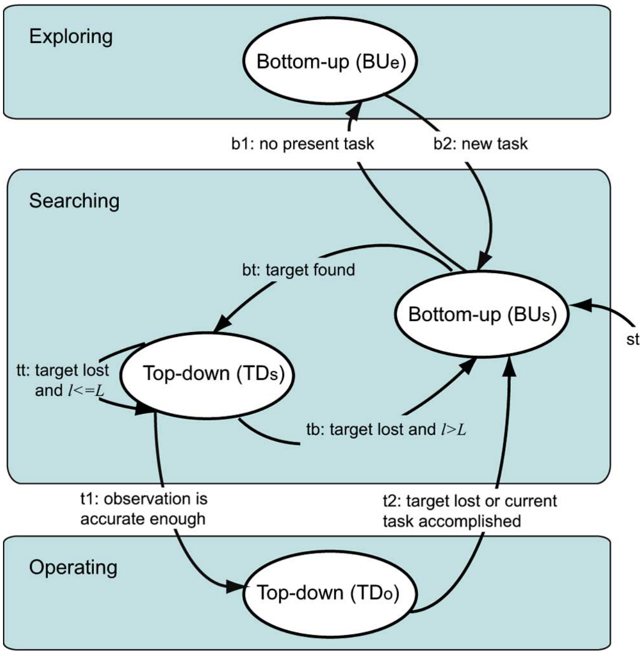

Fig. 2. Finite-state machine of the autonomous switching mechanism. Three internal robot modes: exploration, searching, and operation. Four attention states: bottom-up state in exploring mode (BU$_e$), bottom-up state in searching mode (BU$_s$), top-down state in searching mode (TD$_s$), and top-down state in operating mode (TD$_o$). The transitions are discussed in Section III-C.

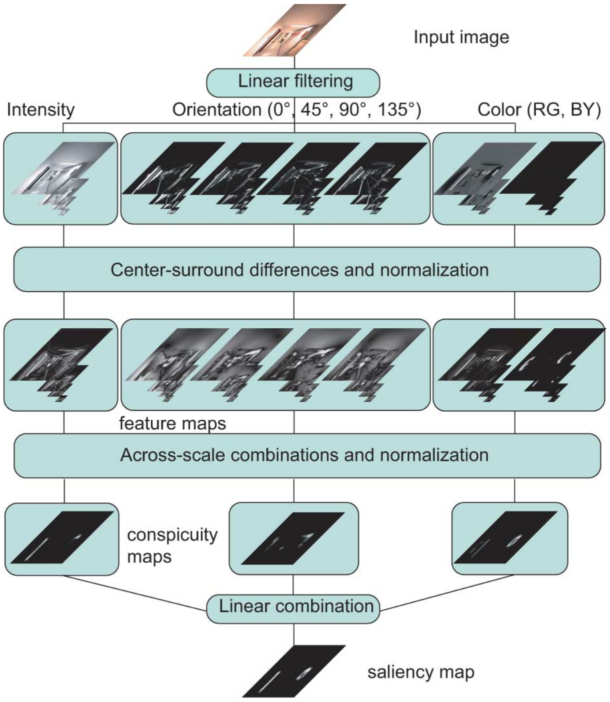
Input image
Linear filtering
Intensity
Orientation (0°, 45°, 90°, 135°)
Color (RG, BY)
Center-surround differences and normalization
feature maps
Across-scale combinations and normalization
conspicuity maps
Linear combination
saliency map

Fig. 3. Saliency-map model.

select candidate regions that may contain target objects. We use a well-known standard computational model for bottom-up attention selection, namely, the saliency map model proposed in [10].

In Fig. 3, the saliency map model is visualized. An input image of, e.g., 640×480 pixels is subsampled into dyadic Gaussian pyramids in three channels (intensity, orientation for $0^\circ, 45^\circ, 90^\circ, 135^\circ$, opponent color in red/green and blue/yellow). The size of the image is reduced from 640×480 to $320 \times 240, \ldots$ 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.

The saliency-map model computes spatial saliency in a static image, which potentially attracts human attention. In technical systems, the saliency map can be used in segmentation of potentially interesting objects from their background. Since dynamic characteristics of the current environment are also an essential cue for robot vision control, temporal novelty in an image sequence should be integrated. Temporal novelty can be modeled as a Bayesian surprise proposed in [40], which is evaluated by the difference between the belief and the perceived information about the world, which in turn measures how novel, surprising, or unexpected the new information is observed. The image region with a higher surprise value is worth being further processed.

Inspired by [40], a local-surprise (LS) map is constructed, which computes the LS value of salient image regions that have already been predicted in the saliency-map computation. LS is defined by applying the Bayesian surprise definition directly on two consecutive saliency maps. The computational details are described as follows. From two consecutive input images without interframe motion, two saliency maps are computed as described in the previous section. Each pixel i with its normalized saliency value λ_i in the saliency map is regarded to be a detector for an LS and is modeled as a probability distribution, which represents the observed saliency value and the observation uncertainty. In [41], Gaussian mixture is used to model the probability distribution. Since Gamma probability density function (pdf) succeeds to model the data input of the local detector in [42], we adopt this modeling and use it to model the belief distribution of the saliency value of a pixel as follows:

$$p_i = \gamma(\lambda_i, \alpha_i, \beta_i) = \frac{\beta_i^{\alpha_i} \lambda_i^{\alpha_i - 1}}{\Gamma(\alpha_i)} e^{-\beta_i \lambda_i} \tag{1}$$

with the shape $\alpha_i > 0$ with an initial value A, the inverse scale $\beta_i > 0$ with an initial value B, and $\Gamma(\cdot)$ the Euler Gamma function.

At time step k, an observation is conducted in which an input image is captured and a respective saliency map is computed. Each pixel in the saliency map is considered to be an LS-detector and obtains a saliency value of $\lambda_{i,k}$. Because of the new data input, the belief or observation of each detector is changed, in which the parameters α_i and β_i evolve as follows [42]:

$$\alpha_{i,k} = A + N_{1\mathrm{D}}(\lambda_{i,k}) \quad \text{and} \quad \beta_{i,k} = \xi \cdot B + 1 \tag{2}$$

where $N_{1\mathrm{D}}$ is a 1-D normalization that scales the saliency value to the range $[0, 1]$, and $\xi \in (0, 1)$ is a forgetting factor.

At the next time step $k + 1$, before a new observation is conducted, it is assumed that the environment does not change. The belief is based on the observation at time step k. The sensed information at each detector or each image pixel is the prior belief distribution and is formulated as follows:

$$p_{i,k} = \gamma(\lambda_{i,k}, \alpha_{i,k}, \beta_{i,k}). \tag{3}$$

After a new image is captured at time step $k + 1$, the detector/pixel i has a new saliency value of $\lambda_{i,k+1}$. The belief of the environment is updated, in which the parameters α_i and β_i evolve as follows:

$$\alpha_{i,k+1} = \alpha_{i,k} + N_{1\mathrm{D}}(\lambda_{i,k+1}) \quad \text{and} \quad \beta_{i,k+1} = \xi \cdot \beta_{i,k} + 1. \tag{4}$$

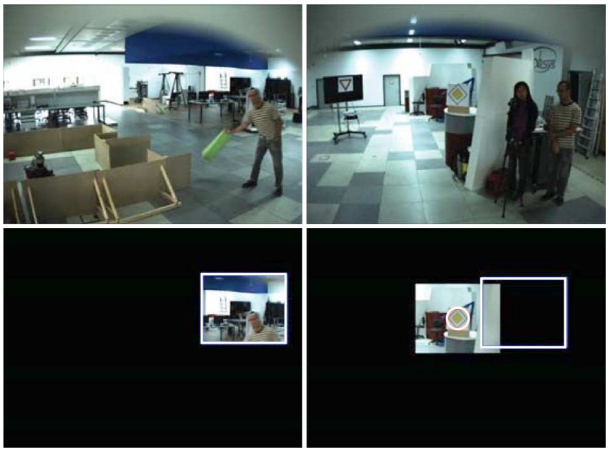

Fig. 4. (Left) Bottom-up state. (Right) Top-down state. (Top) Original input images. (Bottom) Resultant robot attention windows. (Rectangles) Salient/surprising image regions (the same as the masked region in bottom-up state). (Masked regions). Current robot FOA. (Circle) Detected target object.

Then, the posterior belief distribution can be formulated as follows:

$$p_{i,k+1} = \gamma(\lambda_{i,k+1}, \alpha_{i,k+1}, \beta_{i,k+1}). \tag{5}$$

To quantify the distance of the prior belief and the posterior belief of a detector about the sensed environment, the LS for the pixel i is defined as Kullback–Leiber divergence $\tau(x,y)$ (also the relative entropy) between the posterior and prior saliency distributions, which is formulated as follows:

$$\begin{aligned}\tau_{i,k+1} &= \mathrm{KL}(p_{i,k+1} \| p_{i,k})\\ &= -\alpha_{i,k} + \alpha_{i,k} \log \frac{\beta_{i,k+1}}{\beta_{i,k}} + \log \frac{\Gamma(\alpha_{i,k})}{\Gamma(\alpha_{i,k+1})}\\ &\quad + \beta_{i,k} \frac{\alpha_{i,k+1}}{\beta_{i,k+1}} + (\alpha_{i,k+1} - \alpha_{i,k})\Psi(\alpha_{i,k+1}) \text{ in [bit]}\end{aligned} \tag{6}$$

where $\Psi(\cdot)$ is the Digamma function.

The LS is an indicator of a bottom-up robot gaze control. It combines the temporal novelty and the spatial saliency in a way that the larger the interframe saliency variation of the pixel i is and the higher the saliency value the pixel i contains, the higher the LS value $\tau_{i,k+1}$ is. Compared with [42], which aims at a best explanation of human behavior in an active search for nonspecific information of subjective interest, the surprise concept is used here for detection of onset, offset, and motion of salient foreground objects, which focuses on robot applications. The initial values of the distribution parameters α and β are kept constant to emphasize saliency variation at the current time step and enhance the sensitivity of awareness of dynamic environments. Another metric "global surprise" is defined in another work of ours [43] to describe the global dynamics of the environment and resemble the damping effect. This way, local surprise and global surprise are independent of each other and can be flexibly applied alone or together in different contexts.

In the example shown in the left column in Fig. 4, the rectangles in solid lines are the FOA predicted by the surprise map. A moving human is selected as the FOA because of its high surprise value. In the 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 FOA (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 [44]. In the bottom-up state, the salient/surprising image regions in an input image are viewed sequentially according to their descending saliency/surprise values.

The difference between the visual behaviors in the state BU_s and BU_e is whether exhaustive search is applied in the selected FOA. In the BU_s state, the object detection algorithm is applied in the selected FOA, since the robot has a specific task in the searching mode. In the BU_e state, the robot only attends toward the salient/surprising region. No further information processing is applied here.

B. Top-Down State

In the top-down state, the robot concentrates on the image region that contains 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 that contains the target object to get better accuracy.

Fig. 4 right shows an example of the FOA selection in the top-down state. A robot is supposed to detect a traffic sign and approach it. The region around a target object, i.e., the masked region in the right-bottom image, is selected as the current robot FOA and is further processed in detail, although this region is not the most salient/surprising region at this moment; this is, in fact, the region in the rectangle.

In short, in the top-down state, the previous position of the target object is known. No matter how salient/surprising the other features are, to perform its task, the robot attends to the target object.

The difference between the behaviors in the state TD_s and TD_o is that in the TD_s state, the observation of the target object has a higher priority, while in state TD_o, the robot starts to accomplish its task based on the complete observation acquired in TD_s state.

C. Switching Mechanism

The main contribution of this section is to realize autonomous switching between the top-down and the bottom-up visual attention selections that consider robot task performance. The transition conditions illustrated in Fig. 2 are defined as follows.

After initialization, the image region to be further processed is selected in the BU_s state, since the position of the target object is unknown at this moment. Once a target is found in the selected FOA, the TD_s state is activated (bt). In this state, the image region around the target is selected constantly, while the other salient features are ignored. If the target is lost, e.g., because of lighting condition change or humans and vehicles that hide the target object, the robot should initially continue focusing on the last region for L frames to see if the target object is redetectable (tt). If the robot stays in top-down state for l frames, i.e., $l > L$, and the target is still unseen, the BU_s state is triggered again to search for the previous target (tb).

If the observation of the target object in the TD_s state is accurate enough, the robot starts to operate ($t1$). To evaluate the observation uncertainty, the n-D system state $\underline{x} \in \mathbb{R}^n$ of the current robot task is modeled as an n-D Gaussian distribution with mean value $\underline{\mu}$ and covariance matrix $R_{\underline{x}}$ in the task space computed using a Kalman filter (KF). The system state $\underline{x}$ is chosen according to the current task and can be the 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 $k-1$ is regarded as the prior pdf p_{k-1}, while the posterior belief distribution about the system state at the current time step k is p_k with a continuous variable $\underline{x}$ for specific tasks, which is

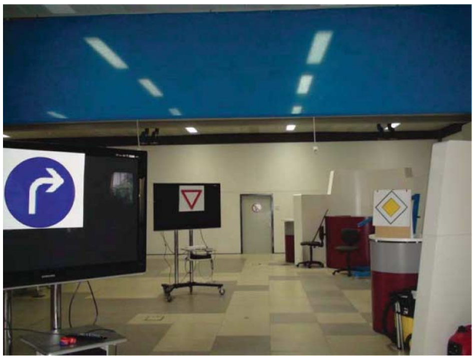

Fig. 5. Experimental setup consisting of three different signs.

defined as follows:

$$p_k = \frac{1}{(\sqrt{2\pi})^n (\det R_{\underline{x},k})^{1/2}} \times \exp\left(-\frac{1}{2}\left(\underline{x}_k - \underline{\mu}_k\right)^T \left(R_{\underline{x},k}\right)^{-1}\left(\underline{x}_k - \underline{\mu}_k\right)\right). \quad (7)$$

Then, the observation uncertainty is defined as the Kullback–Leibler divergence or relative entropy computed as follows:

$$\mathrm{KL}(p_k \| p_{k-1}) = \int_{-\infty}^{\infty} p_k \log \frac{p_k}{p_{k-1}} \mathrm{d}\underline{x} \text{ in [bit]}. \quad (8)$$

An empirical threshold is defined for the relative entropy between the predicted and the updated state estimate as one of the criteria to evaluate the observation uncertainty. The smaller the observation uncertainty is, the less the estimation and its expected value vary, and therefore, the more certain the position estimation is. If the observation uncertainty at the kth step is smaller than this threshold, the observation at this step is regarded as successfully executed. Upon this value, the robot takes the decision for what action to be performed next: operating or observing. Correspondingly, if the task is finished or the target is lost, the robot stops the current operation, turns into the BU_s state, and observes $(t2)$.

If the predefined task is accomplished in total, the robot explores the world by directing its attention toward interesting areas in the environment selected in a pure bottom-up state BU_e $(b1)$. If a new task with new target objects arrives, the robot attention selection is in the BU_s state again $(b2)$.

To sum up, the robot's visual behavior with different emphases of information acquisition is now adapted to the internal robot state. The switching of robot FOA selection mechanisms is autonomously conducted.

IV. Performance Evaluation

To demonstrate the strategy, experiments were conducted using the ACE robot. Sign detection and approaching tasks were assigned to the ACE robot. Fig. 5 shows the experimental scenario in the institute laboratory from the robot perspective. Three different signs were placed in different distances to the initial robot position.

These signs cannot be uniquely described by low-level features used in the saliency-map model and, therefore, cannot be easily recognized and distinguished by enhancing certain bottom-up features using top-down information. Previously trained classifiers based on Haar-like features are used for object recognition [45]. To lower the computational cost of object recognition, the classifiers were only applied in the FOA selected in the input images. The whole input image represents a peripheral sensor input, while the focus region represents a foveated sensor input with a higher resolution.

A. Experiment 1: Searching → Operating → Exploring

In the first experiment, the robot was supposed to detect the blue sign and move toward this sign. If the robot reached its desired position, namely, 1 m in front of the sign, it should turn 180° and move back to its initial position.

Fig. 6 illustrates some representative original input images and their respective results of the attention selection (the masked region). The frame number and the attention selection state of each image are also given. The robot first searched for the traffic sign in the BU_s state. The object detection algorithm was applied in the selected salient/surprising image regions (frame 20 and 24). After the sign was detected in the FOA in frame 24, the robot kept focusing on the sign and computed the relative position (frame 26) in the TD_s state. After the position estimation of the sign was accurate enough, the robot started to move toward the sign and tracked the sign during the movement (frame 120) in the TD_o state. The threshold for the observation uncertainty was set to 0.12 bit. After the task was accomplished, the robot turned back and looked at the salient/surprising parts in the environment (frames 144, 153, 176, and 177) in the BU_e state. The size of FOA varied with the size of the detected target object or the size of the salient/surprising regions.

B. Experiment 2: Searching ↔ Operating

In the second experiment, 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 1 m in front of the sign. Then, the robot head should turn to another direction and search for the next sign.

Fig. 7 illustrates the experimental results. Images with the FOA (the masked region) and salient/surprising region (the region in solid lines) as well as the frame number are shown. At the first step, ACE looked straight ahead and the BU_s state was activated. In frame 1, the blue sign was detected. The FOA changed into the TD_s state. The image region around the blue sign was selected in the following frames, until the robot reached the position 1 m in front of the blue sign (frame 44). Then, the robot turned its head randomly to the right side and detected the yellow sign (frame 45). 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 the BU_s state was activated after several frames. In frame 127, the yellow sign was redetected in the FOA. The TD_s state was triggered again. After the robot reached the position 1 m in front of the yellow sign, the head was randomly directed and the state was the BU_s state again (frame 149). In frames 151 and 214, the red sign was detected and tracked. For 228 frames in total, there are 18 frames in the bottom-up state and 210 frames in the top-down state.

Fig. 8 illustrates the evolution of the observation uncertainty and the switching mechanism between the top-down and the bottom-up states. The semitransparent time intervals indicate the operating state in which the robot was moving. The blank areas indicate the time intervals in which the robot was in the searching mode. The frame numbers near the arrows show several representative time points. In frame 1, the first sign was detected. The observation uncertainty reached its maximum in frame 4 and decreased in the searching mode, since repeated viewing of the same object reduced the observation uncertainty. In frame 36, the observation uncertainty reached its threshold, here 0.12 bit. The robot was triggered into the operating mode and started to move toward the first sign. In frame 45, the second sign was detected coincidently in the top-down search window of the first sign in the operating mode.

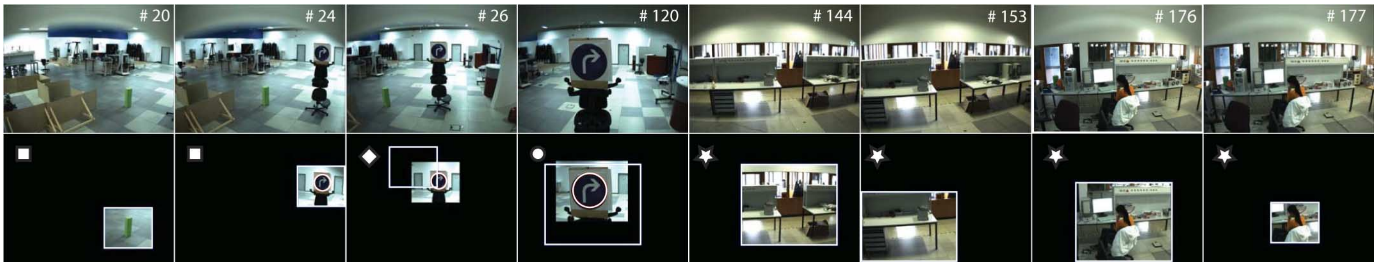

Fig. 6. Results of experiment 1 comprising original input images (the first row) and their respective resultant images with the robot FOA (the second rows). Numbers on the original images indicate the frame number. The markers on the resultant images are defined as follows: (Rectangle) Bottom-up attention in searching state (BU_s). (Diamond) Top-down attention in searching state (TD_s). (Circle) Top-down attention in operating state (TD_o). (Star) Bottom-up attention in exploring state (BU_e). (Rectangles) Salient/surprising image regions. (Masked regions) Current robot FOA. (Circle) Detected target object.

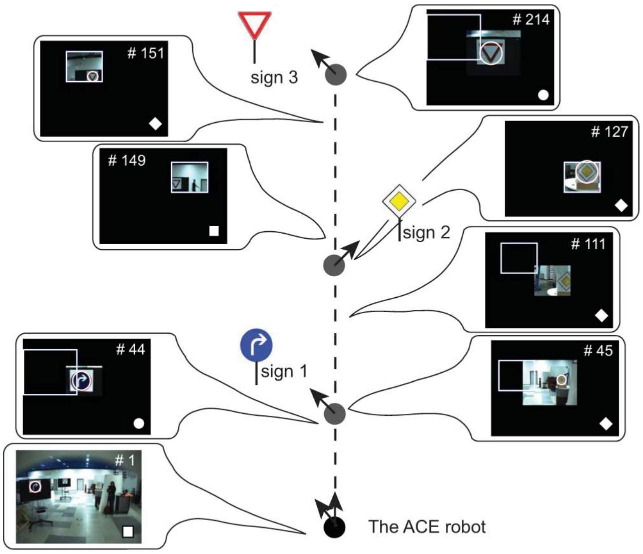

Fig. 7. Results of experiment 2 comprising the resultant images of robot FOA. The same markers are used in Fig. 6.

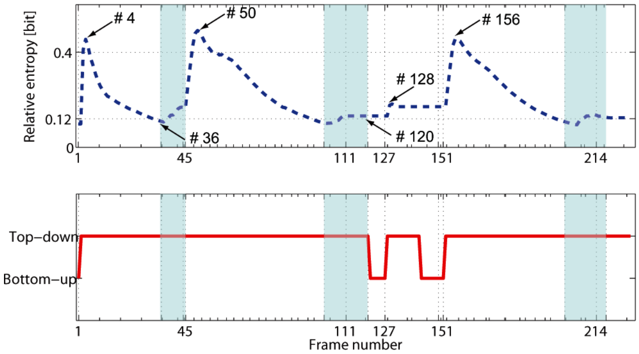

Fig. 8. Relative entropy (observation uncertainty) evolution (dashed line) and the respective attention state (solid line). The semitransparent areas indicate the time intervals in which the robot was in operating state. The blank areas indicate the time intervals in which the robot was in searching state. Some representative time points are shown with their frame numbers.

TABLE I
AVERAGE COMPUTATION TIME IN THE EXPERIMENT

Task	Time [ms]
Image capture (approximately)	67
Surprise map computation	20
Search for a sign in the FOA	31
Search for 3 signs in the FOA	33
Search for a sign in the whole image	183
Search for 3 signs in the whole image	373

Therefore, the robot attention state was switched from this top-down state to the top-down state in the searching mode for the second sign. The observation uncertainty reached a local maximum in frame 50. The second sign was lost in L frames before frame 120 and redetected in frame 127. A local maximum of the observation uncertainty was reached in frame 128. In frame 156, a local maximum of the observation uncertainty was reached again, after the third sign was detected.

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, a very low threshold value was set to the observation uncertainty, which caused a relatively long period in the searching mode before the robot started to operate. However, this can be easily improved if other sensor modalities are used for obstacle avoidance as well.

Table I shows the average computation time that was taken in different phases. Since the bottom-up attention selection was implemented on a platform of multiple graphics processing units (GPUs), real-time processing in this part is ensured. The most expensive processing is the object recognition using the previously trained classifiers. There is a large improvement in the performance if the robot searches for the signs only in the FOA but not in the whole image.

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 between them. Purely bottom-up attention facilitates the robot task accomplishment by providing FOA candidates and reducing the detection time. Image regions with higher saliency are regarded as positions with a higher probability of containing a target object and are processed first. Moreover, *inhibition-of-return* (IOR) is used here to extend the robot FOA to less salient regions.

In this experiment, the resolution of the vision sensor is still sufficient for sign recognition. If more resolution is required to further process the selected region, the bottom-up state is a must for efficient utilization of high-resolution cameras that provide potential image region candidates before a target object is found. Otherwise, the high-resolution camera has to search for objects in the environment randomly and inefficiently. In addition, the bottom-up state guides the robot attention to explore the environment in a flexible way if top-down information does not exist in the current FOV at all.

To accelerate the whole task performance, three solutions are suggested as follows. 1) Reduce the computation time for the bottom-up state from our previous work using the multi-GPU implementation [46]. 2) Use top-down biased bottom-up attention selection for a more efficient search in the top-down state. 3) Apply IOR in the 3-D task space to avoid repeated view of the positions that have already been observed. Currently, a simple IOR is integrated in our implementation in the way that the current FOA is suppressed in the searching and exploring modes where the robot is not in motion.

V. Conclusion and Future Work

To enhance the ability of bottom-up attention to facilitate robot task performance, autonomous switching between top-down and bottom-up attention selections has been realized, which fills a gap in situations where totally different targets are searched while contexts vary. This capability of autonomous switching of visual attention selection models enables a vision-guided mobile robot to be "autonomous" in this aspect. Visual behavior, which is the selection of attention focus, is autonomously adapted to robot's internal state.

To demonstrate our strategies, the active multifocal camera system placed on the ACE robot was used in the experiments. Thereby, the cooperation of the hardware and the parallel implementation that aims at real-time robot visual attention control during robot motion is established sufficiently. This application-oriented robot attention system makes a step forward in efficient visual information selection and benefits a further development toward cognitive visual perception in the robotics domain.

It is worth to mention that visual attention can support object search but does not guarantee the success. Since the bottom-up attention selection selects salient positions in the environment, some salient but task-irrelevant candidates could also be selected, such as the light spots and the blue wall in Fig. 5. Extensions using context information is envisioned [47].

Acknowledgment

The authors would like to thank Prof. G. Cheng for valuable discussions about this paper and the other Autonomous City Explorer (ACE)-Team members for their excellent work to design and implement the ACE platform.

References

[1] T. Xu, H. Wu, T. Zhang, K. Kühnlenz, and M. Buss, "Environment adapted active multi-focal vision system for object detection," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2009, pp. 2418–2423.

[2] A. Bauer, K. Klasing, G. Lidoris, M. 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," *Int. J. Soc. Robot.*, vol. 1, no. 2, pp. 127–140, 2009.

[3] K. Kühnlenz, M. Bachmayer, and M. Buss, "A multi-focal high-performance vision system," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2006, pp. 150–155.

[4] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cogn. Psychol.*, vol. 12, pp. 97–136, 1980.

[5] J. Duncan and G. W. Humphreys, "Visual search and stimulus similarity," *Psychol. Rev.*, vol. 96, pp. 433–458, 1989.

[6] J. M. Wolfe, "Guided search 2.0: A revised model of visual search," *Psychonomic Bull. Rev.*, vol. I, no. 2, pp. 202–238, 1994.

[7] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Hum. Neurobiol.*, vol. 4, pp. 219–227, 1985.

[8] R. Milanese, H. Wechsler, S. Gil, J. M. Bost, and T. Pun, "Integration of bottom-up and top-down cues for visual attention using non-linear relaxation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 1994, pp. 781–785.

[9] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artif. Intell.*, vol. 78, pp. 507–545, 1995.

[10] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *Pattern Anal. Mach. Intell.*, vol. 20, pp. 1254–1259, 1998.

[11] Y. Sun and R. Fisher, "Object based visual attention for computer vision," *Artif. Intell.*, vol. 146, no. 1, pp. 77–123, 2003.

[12] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: A Bayesian framework for saliency using natural statistics," *J. Vis.*, vol. 8, no. 7, pp. 1–20, 2008.

[13] N. D. B. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *J. Vis.*, vol. 9, no. 3, pp. 1–24, 2009.

[14] D. Walther, U. Rutishauser, C. Koch, and P. Perona, "Selective visual attention enables learning and recognition of multiple objects in cluttered scenes," *Comput. Vis. Image Understanding*, vol. 100, pp. 41–63, 2005.

[15] B. A. Draper and A. Lionelle, "Evaluation of selective attention under similarity transformations," *Comput. Vis. Image Understanding, Spec. Issue: Attention Perform. Comput. Vis.*, vol. 100, no. 1–2, pp. 152–171, 2005.

[16] V. Navalpakkam and L. Itti, "An integrated model of top-down and bottom-up attention for optimizing detection speed," in *Proc. IEEE Comp. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, vol. 2, pp. 2049–2056.

[17] G. Fritz, C. Seifert, L. Paletta, and H. Bischof, "Attentive object detection using an information theoretic saliency measure," in *Attention and Performance in Computational Vision*, vol. 3368/2005 (Lecture Notes in Computer Science). New York: Springer-Verlag, 2005, pp. 29–41.

[18] J. Han, K. N. Ngan, M. Li, and H. Zhang, "Unsupervised extraction of visual attention objects in color images," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 16, no. 1, pp. 141–145, Jan. 2006.

[19] N. Ouerhani and H. Hügli, "A model of dynamic visual attention for object tracking in natural image sequences," in *Computational Methods in Neural Modeling* (Lecture Notes in Computer Science). New York: Springer-Verlag, 2003, pp. 702–709.

[20] G. Backer and B. Mertsching, "Two selection stages provide efficient object-based attentional control for dynamic vision," in *Proc. Int. Workshop Attention Perform. Comput. Vis.*, 2003, pp. 9–16.

[21] S. Frintrop and M. Kessel, "Most salient region tracking," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2009, pp. 1869–1874.

[22] G. Heidemann, R. Rae, H. Bekel, I. Bax, and H. Ritter, "Integrating context-free context-dependent attentional mechanisms for gestural object reference," *Computer Vision Systems*, vol. 2626/2003 (Lecture Notes in Computer Science). New York: Springer-Verlag, pp. 22–33, 2003.

[23] B. Khadhouri and Y. Demiris, "Compound effects of top-down and bottom-up influences on visual attention during action recognition," in *Proc. 19th Int. Jont. Conf. Artif. Intell.*, 2005, pp. 1458–1463.

[24] J. Peng, A. Peters, X. Ao, and A. Srikaew, "Grasping a waving object for a humanoid robot using a biologically-inspired active vision system," in *Proc. IEEE Int. Workshop Robot Human Interactive Commun.*, 2003, pp. 115–120.

[25] S. Vijayakumar, J. Conradt, T. Shibata, and S. Schaal, "Overt visual attention for a humanoid robot," in *Proc. Int. Conf. Intell. Robots Syst.*, 2001, vol. 4, pp. 2332–2337.

[26] C. Scheier and S. Egner, "Visual attention in a mobile robot," in *Proc. IEEE Int. Symp. Ind. Electron.*, 1997, vol. 1, pp. SS48–SS52.

[27] S. Schaal and L. Itti, "Learning and attention with a humanoid robot head," Nat. Sci. Foundation (NSF) Exhibition, 2005.

[28] L. Itti, N. Dhavale, and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," presented at the 48th Annu. Int. Symp. Optical Sci. Technol., San Diego, CA, 2003.

[29] A. Ude, V. Wyart, L. H. Lin, and G. Cheng, "Distributed visual attention on a humanoid robot," in *Proc. 5th IEEE-RAS Int. Conf. Humanoid Robots*, 2005, pp. 381–386.

[30] S. W. Ban and M. Lee, "Biologically motivated vergence control system based on stereo saliency map model," in *Scene Reconstruction, Pose Estimation and Tracking*. Hong Kong: I-Tech, 2007, pp. 513–530.

[31] M. Björkman and J. O. Eklundh, "Vision in the real world: Attending and recognizing objects," *Int. J. Imag. Syst. Technol.*, vol. 16, pp. 189–208, 2007.

[32] S. Frintrop and P. Jensfelt, "Active gaze control for attentional visual SLAM," in *Proc. IEEE Int. Conf. Rob. Autom.*, 2008, pp. 3690–3697.

[33] B. Rasolzadeh, A. Tavakoli, and J.-O. Eklundh, "An attentional system combining top-down and bottom-up influences," in *Proc. Workshop Attention Perform. Comput. Vis.*, 2007, pp. 123–140.

[34] J. K. Tsotsos and K. Shubina, "Attention and visual search: Active robotic vision systems that search," in *Proc. 5th Int. Conf. Comput. Vis. Syst.*, 2007.

[35] P. E. Forssen, D. Meger, K. Lai, S. Helmer, J. J. Little, and D. G. Lowe, "Informed visual search: Combining attention and object recognition," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2008, pp. 935–942.

[36] S. Frintrop, "Vocus: A visual attention system for object detection and goal-directed search," Ph.D. dissertation, Inst. Comput. Sci., Rheinische Friedrich-Wilhelms-Univ. Bonn, Bonn, Germany, 2005.

[37] C. Breazeal and B. Scassellati, "A context-dependent attention system for a social robot," in *Proc. 16th Int. Joint. Conf. Artif. Intell.*, 1999, pp. 1146–1153.

[38] Y. Yu, G. K. I. Mann, and R. G. Gosine, "A task-driven object-based attention model for robots," in *Proc. IEEE Int. Conf. Robot. Biomimetics*, 2007, pp. 1751–1756.

[39] T. Xu, N. Chenkov, K. Kühnlenz, and M. Buss, "Autonomous switching of top-down and bottom-up attention selection for vision guided mobile robots," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2009, pp. 4009–4014.

[40] L. Itti and P. Baldi, "A principled approach to detecting surprising events in video," in *Proc. IEEE Comp. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 631–637.

[41] G. Boccignone, "Nonparametric Bayesian attentive video analysis," in *Proc. 19th Int. Conf. Pattern Recognit.*, 2008, pp. 1–4.

[42] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vis. Res.*, vol. 49, no. 10, pp. 1295–1306, 2009.

[43] T. Xu, "Aspects of visual attention for autonomous mobile robots," Ph.D. dissertation, Technische Univ. München, München, Germany, 2010.

[44] T. Xu, Q. Mühlbauer, S. Sosnowski, K. Kühnlenz, and M. Buss, "Looking at the surprise: Bottom-up attention control of an active camera system," in *Proc. 10th Int. Conf. Control, Autom., Robot. Vis.*, 2008, pp. 637–642.

[45] Q. Mühlbauer, S. Sosnowski, T. Xu, T. Zhang, and K. Kühnlenz, M. Buss, "Navigation through urban environments by visual perception and interaction," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2009, pp. 1907–1913.

[46] 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. Robot. Autom.*, 2009, pp. 41–47.

[47] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search," *Psychol. Rev.*, vol. 113, no. 4, pp. 766–786, 2006.

Trajectory Planning of Unicycle Mobile Robots With a Trapezoidal-Velocity Constraint

M. Haddad, W. Khalil, and H. E. Lehtihet

***Abstract*—We propose an efficient stochastic scheme for minimum-time trajectory planning of a nonholonomic unicycle mobile robot under constraints on path curvature, velocities, and torques. This problem, which is known to be complex, often requires important runtimes, particularly if obstacles are present and if full dynamics is considered. The proposed technique is a fast variant of the random-profile approach recently applied to wheeled-mobile robots. It incorporates a trapezoidal-velocity-profile constraint that helps reduce the number of unknown parameters and that speeds up the calculation steps. Results are presented for two- and three-wheel mobile robots in free/constrained workspaces. A comparison with reference solutions, which were obtained independently, shows that the proposed variant is able to achieve almost the same quality of calculated trajectories while reducing the runtime considerably.**

***Index Terms*—Stochastic optimization, trajectory planning, trapezoidal-velocity profile (TVP), wheeled-mobile robot.**

I. Introduction

There has been a growing interest in the development of robotized systems, such as fixed-base manipulators (FBMs), wheeled-mobile platforms (WMPs), and wheeled-mobile manipulators (WMMs). Because of their potentialities, these systems have become an essential tool to execute complex tasks. The fields of applications are manifold: inspection/control operations, manipulation of radioactive/toxic materials, exploration, etc. Some of these applications require the use of a trajectory planner that yields, for a given performance criterion, optimal or near-optimal solutions while considering full dynamics.

Techniques that are able to achieve this goal can be grouped in two categories [1]: *direct* and *indirect* methods. These latter use Pontryagin's maximum principle (PMP) [2] and write optimality conditions as a boundary-value problem [3]–[5]. The former use discretization of state/control variables and convert the trajectory problem to a parametric optimization, which is solved via nonlinear programming [1] or stochastic means [6], [7].

The random-profile approach (RPA) is one of the *direct* methods that is able to account for full dynamics. It can be applied to various systems (see, e.g., FBM [8], WMP [9], and WMM [10]). It is characterized by some interesting features, which are as follows.

1) *Versatility*: It can solve diverse types of trajectory problems (generalized/operational point-to-point tasks with free/constrained paths in free/constrained workspaces) using diverse types of performance criteria (time of travel, energy, etc.) under diverse types of kinodynamics constraints (bounded velocities/torques/curvature, stability, etc.) [8]–[12].

Manuscript received November 23, 2009; revised May 2, 2010; accepted July 21, 2010. Date of publication September 7, 2010; date of current version September 27, 2010. This paper was recommended for publication by Associate Editor T. Murphey and Editor J.-P. Laumond upon evaluation of the reviewers' comments.

M. Haddad and H. E. Lehtihet are with the Laboratory of Structure Mechanics, Ecole Militaire Polytechnique (EMP), 16111 Algiers, Algeria (e-mail: haddadmoussa2003@yahoo.fr; he.lehtihet@gmail.com).

W. Khalil is with the Institut de Recherche en Communications et Cybernétique de Nantes—UMR Centre National de la Recherche Scientifique 6597, Ecole Centrale de Nantes, 44321 Nantes, France (e-mail: wisama.khalil@irccyn.ec-nantes.fr).

Digital Object Identifier 10.1109/TRO.2010.2062090