2

Development of an Autonomous Visual Perception System for Robots Using Object-Based Visual Attention

Yuanlong Yu, George K. I. Mann and Raymond G. Gosine
Faculty of Engineering and Applied Science, Memorial University of Newfoundland
St. John's, NL,
Canada

1. Introduction

Unlike the traditional robotic systems in which the perceptual behaviors are manually designed by programmers for a given task and environment, autonomous perception of the world is one of the challenging issues in the cognitive robotics. It is known that the *selective attention* mechanism serves to link the processes of perception, action and learning (Grossberg, 2007; Tipper et al., 1998). It endows humans with the *cognitive capability* that allows them to *learn* and *think* about how to perceive the environment autonomously. This visual attention based autonomous perception mechanism involves two aspects: *conscious aspect* that directs perception based on the current task and learned knowledge, and *unconscious aspect* that directs perception in the case of facing an unexpected or unusual situation. The *top-down attention* mechanism (Wolfe, 1994) is responsible for the conscious aspect whereas the *bottom-up attention* mechanism (Treisman & Gelade, 1980) corresponds to the unconscious aspect. This paper therefore discusses about how to build an artificial system of autonomous visual perception.

Three fundamental problems are addressed in this paper. The first problem is about pre-attentive segmentation for object-based attention. It is known that attentional selection is either *space-based* or *object-based* (Scholl, 2001). The space-based theory holds that attention is allocated to a spatial location (Posner et al., 1980). The object-based theory, however, posits that some pre-attentive processes serve to segment the field into discrete objects, followed by the attention that deals with one object at a time (Duncan, 1984). This paper proposes that object-based attention has the following three advantages in terms of computations: 1) Object-based attention is more robust than space-based attention since the attentional activation at the object level is estimated by accumulating contributions of all components within that object, 2) attending to an exact object can provide more useful information (e.g., shape and size) to produce the appropriate actions than attending to a spatial location, and 3) the discrete objects obtained by pre-attentive segmentation are required in the case that a global feature (e.g., shape) is selected to guide the top-down attention. Thus this paper adopts the object-based visual attention theory (Duncan, 1984; Scholl, 2001).

Although a few object-based visual attention models have been proposed, such as (Sun, 2008; Sun & Fisher, 2003), developing a pre-attentive segmentation algorithm is still a challenging issue as it is a unsupervised process. This issue includes three types of challenges: 1) The

ability to automatically determine the number of segments (termed as *self-determination*), 2) the computational efficiency, and 3) the robustness to noise. Although K-labeling methods (e.g., normalized cut (Shi & Malik, 2000)) can provide the accuracy and robustness, they are ineffective and inefficient when the number of segments is unknown. In contrast, recent split-and-merge methods (e.g., irregular pyramid based segmentation (Sharon et al., 2006)) are capable of determining the number of segments and computationally efficient, whereas they are not robust to noise. This paper proposes a new pre-attentive segmentation algorithm based on the irregular pyramid technique in order to achieve the self-determination and robustness as well as keep the balance between the accuracy and efficiency.

The second problem is about how to model the attentional selection, i.e., model the cognitive capability of *thinking* about what should be perceived. Compared with the well-developed bottom-up attention models (Itti & Baldi, 2009; Itti et al., 1998), modeling the top-down attention is far from being well-studied. The top-down attention consists of two components: 1) Deduction of task-relevant object given the task and 2) top-down biasing that guides the focus of attention (FOA) to the task-relevant object. Although some top-down methods have been proposed, such as (Navalpakkam & Itti, 2005), several challenging issues require further concerns. Since the first component is greatly dependent on the knowledge representation, it will be discussed in the next paragraph. Regarding the second component, the first issue is about the effectiveness of top-down biasing. The main factor that decays the effectiveness is that the task-relevant object shares some features with the distracters. It indicates that the top-down biasing method should include a mechanism to make sure that the task-relevant object can be discriminated from distracters. The second issue is about the computational efficiency based on the fact that the attention is a fast process to select an object of interest from the image input. Thus it is reasonable to use some low-level features rather than high-level features (e.g., the iconic representation (Rao & Ballard, 1995a)) for top-down biasing. The third one is the adaptivity to automatically determine which feature(s) is used for top-down biasing such that the requirement of manually re-selecting the features for different tasks and environment is eliminated. This paper attempts to address the above issues by using the integrated competition (IC) hypothesis (Duncan et al., 1997) since it not only summarizes a theory of the top-down attention, which can lead to a computational model with effectiveness, efficiency and adaptivity, but also integrates the object-based attention theory. Furthermore, it is known that bottom-up attention and top-down attention work together to decide the attentional selection, but how to combine them is another challenging issue due to the multi-modality of bottom-up saliency and top-down biases. A promising approach to this issue is setting up a unified scale at which they can be combined.

The third problem is about the cognitive capability of autonomously *learning* the knowledge that is used to guide the conscious perceptual behavior. According to the psychological concept, the memory used to store this type of knowledge is called long-term memory (LTM). Regarding this problem, the following four issues are addressed in this paper. The first issue is about the unit of knowledge representations. Object-based vision theory (Duncan, 1984; Scholl, 2001) indicates that a general way of organizing the visual scene is to parcel it into discrete objects, on which perception, action and learning perform. In other words, the internal attentional representations are in the form of objects. Therefore objects are used as the units of the learned knowledge. The second issue is what types of knowledge should be modeled for guiding the conscious perceptual behavior. According to the requirements of the attention mechanism, this paper proposes that the knowledge mainly includes LTM task representations and LTM object representations. The *LTM task representation* embodies the

association between the attended object at the last time and predicted task-relevant object at the current time. In other words, it tells the robot what should be perceived at each time. Thus its objective is to deduce the task-relevant object given the task in the attentional selection stage. The *LTM object representation* embodies the properties of an object. It has two objectives: 1) Directing the top-down biasing given the task-relevant object and 2) directing the post-attentive perception and action selection. The third issue is about how to build their structure in order to realize the objectives of these two representations. This paper employs the *connectionist approach* to model both representations as the self-organization can be more effectively achieved by using the cluster-based structure, although some symbolic approaches (Navalpakkam & Itti, 2005) have been proposed for task representations. The last issue is about how to learn both representations through the duration from an infant robot to a mature one. It indicates that a dynamic, constructive learning algorithm is required to achieve the self-organization, such as generation of new patterns and re-organization of existing patterns. Since this paper focuses on the perception process, only the learning of LTM object representations is presented.

The remainder of this paper is organized as follows. Some related work of modeling visual attention and its applications in robotic perception are reviewed in section 2. The framework of the proposed autonomous visual perception system is given in section 3. Three stages of this proposed system are presented in section 4, section 5 and section 6 respectively. Experimental results are finally given in section 7.

2. Related work

There are mainly four psychological theories of visual attention, which are the basis of computational modeling. Feature integration theory (FIT) (Treisman & Gelade, 1980) is widely used for explaining the space-based bottom-up attention. The FIT asserts that the visual scene is initially coded along a variety of feature dimensions, then attention competition performs in a location-based serial fashion by combining all features spatially, and focal attention finally provides a way to integrate the initially separated features into a whole object. Guided search model (GSM) (Wolfe, 1994) was further proposed to model the space-based top-down attention mechanism in conjunction with bottom-up attention. The GSM posits that the top-down request for a given feature will activate the locations that might contain the given feature. Unlike FIT and GSM, the biased competition (BC) hypothesis (Desimone & Duncan, 1995) asserts that attentional selection, regardless of being space-based or object-based, is a biased competition process. Competition is biased in part by the bottom-up mechanism that favors a local inhomogeneity in the spatial and temporal context and in part by the top-down mechanism that favors items relative to the current task. By extending the BC hypothesis, the IC hypothesis (Duncan, 1998; Duncan et al., 1997) was further presented to explain the object-based attention mechanism. The IC hypothesis holds that any property of an object can be used as a task-relevant feature to guide the top-down attention and the whole object can be attended once the task-relevant feature successfully captures the attention.

A variety of computational models of space-based attention for computer vision have been proposed. A space-based bottom-up attention model was first built in (Itti et al., 1998). The surprise mechanism (Itti & Baldi, 2009; Maier & Steinbach, 2010) was further proposed to model the bottom-up attention in terms of both spatial and temporal context. Itti's model was further extended in (Navalpakkam & Itti, 2005) by modeling the top-down attention mechanism. One contribution of Navalpakkam's model is the symbolic knowledge

representations that are used to deduce the task-relevant entities for top-down attention. The other contribution is the multi-scale object representations that are used to bias attentional selection. However, this top-down biasing method might be ineffective in the case that environment contains distracters which share one or some features with the target. Another model that selectively tunes the visual processing networks by a top-down hierarchy of winner-take-all processes was also proposed in (Tsotsos et al., 1995). Some template matching methods such as (Rao et al., 2002), and neural networks based methods, such as (Baluja & Pomerleau, 1997; Hoya, 2004), were also presented for modeling top-down biasing. Recently an interesting computational method that models attention as a Bayesian inference process was reported in (Chikkerur et al., 2010). Some space-based attention model for robots was further proposed in (Belardinelli & Pirri, 2006; Belardinelli et al., 2006; Frintrop, 2005) by integrating both bottom-up and top-down attention.

Above computational models direct attention to a spatial location rather than a perceptual object. An alternative, which draws attention to an object, has been proposed by (Sun & Fisher, 2003). It presents a computational method for grouping-based saliency and a hierarchical framework for attentional selection at different perceptual levels (e.g. a point, a region or an object). Since the pre-attentive segmentation is manually achieved in the original work, Sun's model was further improved in (Sun, 2008) by integrating an automatic segmentation algorithm. Some object-based visual attention models (Aziz et al., 2006; Orabona et al., 2005) have also been presented. However, the top-down attention is not fully achieved in these existing object-based models, e.g., how to get the task-relevant feature is not realized.

Visual attention has been applied in several robotic tasks, such as object recognition (Walther et al., 2004), object tracking (Frintrop & Kessel, 2009), simultaneous localization and mapping (SLAM) (Frintrop & Jensfelt, 2008) and exploration of unknown environment (Carbone et al., 2008). A few general visual perception models (Backer et al., 2001; Breazeal et al., 2001) are also presented by using visual attention. Furthermore, some research (Grossberg, 2005; 2007) has proposed that the adaptive resonance theory (ART) (Carpenter & Grossberg, 2003) can predict the functional link between attention and processes of consciousness, learning, expectation, resonance and synchrony.

3. Framework of the proposed system

The proposed autonomous visual perception system involves three successive stages: pre-attentive processing, attentional selection and post-attentive perception. Fig. 1 illustrates the framework of this proposed system.

Stage 1: The *pre-attentive processing stage* includes two successive steps. The first step is the extraction of pre-attentive features at multiple scales (e.g., nine scales for a 640×480 image). The second step is the pre-attentive segmentation that divides the scene into proto-objects in an unsupervised manner. The *proto-objects* can be defined as uniform regions such that the pixels in the same region are similar. The obtained proto-objects are the fundamental units of attentional selection.

Stage 2: The *attentional selection stage* involves four modules: bottom-up attention, top-down attention, a combination of bottom-up saliency and top-down biases, as well as estimation of proto-object based attentional activation. The bottom-up attention module aims to model the unconscious aspect of the autonomous perception. This module generates a probabilistic location-based bottom-up saliency map. This map shows the conspicuousness of a location compared with others in terms of pre-attentive features. The top-down attention module aims

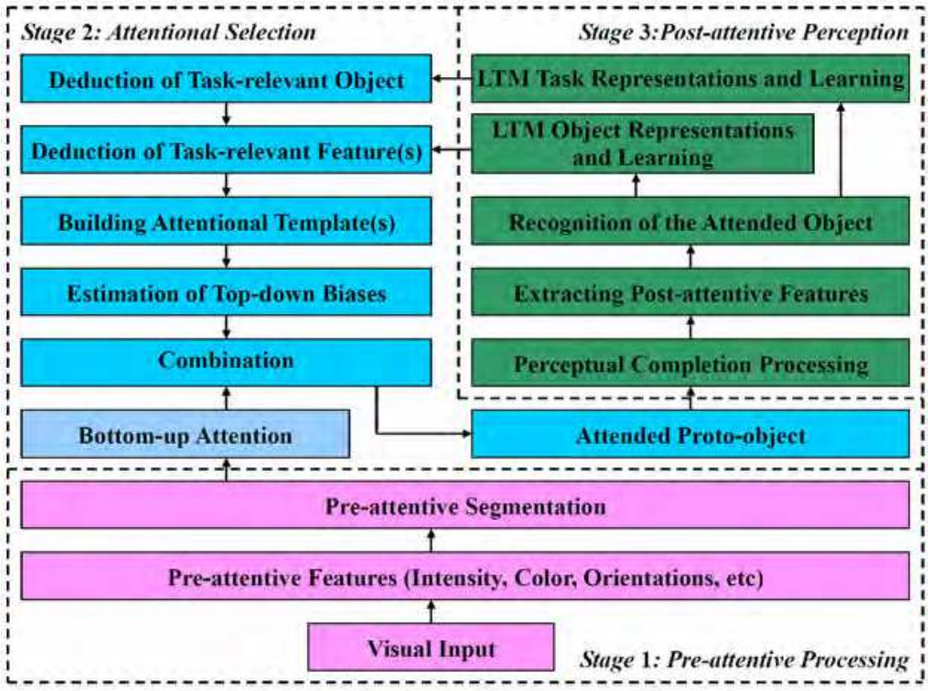


Fig. 1. The framework of the proposed autonomous visual perception system for robots.

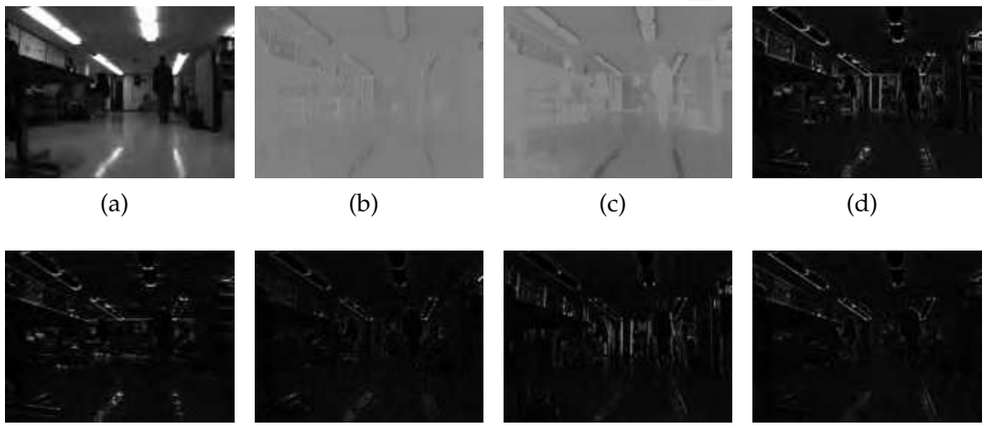
to model the conscious aspect of the autonomous perception. This module is modeled based on the IC hypothesis and consists of four steps. *Step 1* is the deduction of the task-relevant object from the corresponding LTM task representation given the task. *Step 2* is the deduction of the task-relevant feature dimension(s) from the corresponding LTM object representation given the task-relevant object. *Step 3* is to build the attentional template(s) in working memory (WM) by recalling the task-relevant feature(s) from LTM. *Step 4* is to estimate a probabilistic location-based top-down bias map by comparing attentional template(s) with corresponding pre-attentive feature(s). The obtained top-down biases and bottom-up saliency are combined in a probabilistic manner to yield a location-based attentional activation map. By combining location-based attentional activation within each proto-object, a proto-object based attentional activation map is finally achieved, based on which the most active proto-object is selected for attention.

Stage 3: The main objective of the *post-attentive perception stage* is to interpret the attended object in more detail. The detailed interpretation aims to produce the appropriate action and learn the corresponding LTM object representation at the current time as well as to guide the top-down attention at the next time. This paper introduces four modules in this stage. The first module is perceptual completion processing. Since an object is always composed of several parts, this module is required to perceive the complete region of the attended object post-attentively. In the following text, the term *attended object* is used to represent one or all of the proto-objects in the complete region being attended. The second module is the extraction of post-attentive features that are a type of representation of the attended object in WM and used for the following two modules. The third module is object recognition. It functions as a decision unit that determines to which LTM object representation and/or to which instance of that representation the attended object belongs. The fourth module is learning of LTM

object representations. This module aims to develop the corresponding LTM representation of the attended object. The probabilistic neural network (PNN) is used to build the LTM object representation. Meanwhile, a constructive learning algorithm is also proposed. Note that the LTM task representation is another important module in the post-attentive perception stage. Its learning requires the perception-action training pairs, but this paper focuses on the perception process rather than the action selection process. So this module will be discussed in the future work.

4. Pre-attentive processing

4.1 Extraction of pre-attentive features



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 2. Pre-attentive features at the original scale. (a) Intensity. (b) Red-green. (c) Blue-yellow. (d) Contour. (e) - (h) Orientation energy in direction 0°, 45°, 90° and 135° respectively. Brightness represents the energy value.

By using the method in Itti's model (Itti et al., 1998), pre-attentive features are extracted at multiple scales in the following dimensions: intensity $\mathbf{F}_{int}$, red-green $\mathbf{F}_{rg}$, blue-yellow $\mathbf{F}_{by}$, orientation energy $\mathbf{F}_{o_\theta}$ with $\theta \in \{0°, 45°, 90°, 135°\}$, and contour $\mathbf{F}_{ct}$. Symbol $\mathbf{F}$ is used to denote pre-attentive features.

Given 8-bit RGB color components $\mathbf{r}$, $\mathbf{g}$ and $\mathbf{b}$ of the input image, intensity and color pairs at the original scale are extracted as : $\mathbf{F}_{int} = (\mathbf{r}+\mathbf{g}+\mathbf{b})/3$, $\mathbf{F}_{rg} = \mathbf{R}-\mathbf{G}$, $\mathbf{F}_{by} = \mathbf{B}-\mathbf{Y}$, where $\mathbf{R} = \mathbf{r}-(\mathbf{g}+\mathbf{b})/2$, $\mathbf{G} = \mathbf{g}-(\mathbf{r}+\mathbf{b})/2$, $\mathbf{B} = \mathbf{b}-(\mathbf{r}+\mathbf{g})/2$, and $\mathbf{Y} = (\mathbf{r}+\mathbf{g})/2 - |\mathbf{r}-\mathbf{g}|/2 - \mathbf{b}$. Gaussian pyramid (Burt & Adelson, 1983) is used to create the multi-scale intensity and color pairs. The multi-scale orientation energy is extracted using the Gabor pyramid (Greenspan et al., 1994). The contour feature $\mathbf{F}_{ct}(s)$ is approximately estimated by applying a pixel-wise maximum operator over four orientations of orientation energy: $\mathbf{F}_{ct}(s) = \max_{\theta\in\{0°,45°,90°,135°\}} \mathbf{F}_{o_\theta}(s)$, where s denotes the spatial scale. Examples of the extracted pre-attentive features have been shown in Fig. 2.

4.2 Pre-attentive segmentation

This paper proposes a pre-attentive segmentation algorithm by extending the irregular pyramid techniques (Montanvert et al., 1991; Sharon et al., 2000; 2006). As shown in Fig. 3,

the pre-attentive segmentation is modeled as a hierarchical accumulation procedure, in which each level of the irregular pyramid is built by accumulating similar local nodes at the level below. The final proto-objects emerge during this hierarchical accumulation process as they are represented by single nodes at some levels. This accumulation process consists of four procedures.

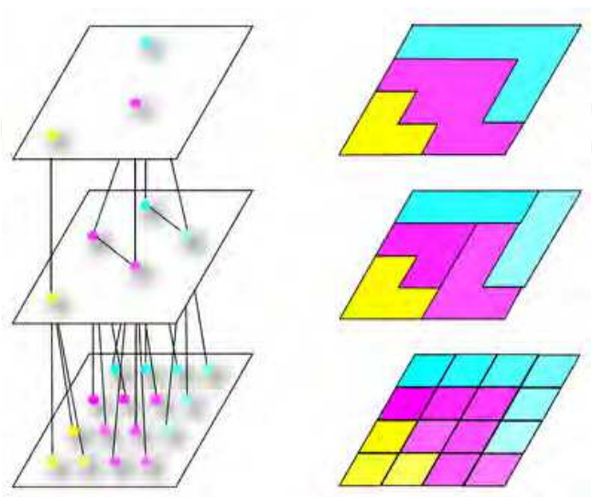


Fig. 3. An illustration of the hierarchical accumulation process in the pre-attentive segmentation. This process is shown from bottom to top. In the left figure, this process is represented by vertices and each circle represents a vertex. In the right figure, this process is represented by image pixels and each block represents an image pixel. The color of each vertex and block represents the feature value. It can be seen that the image is partitioned into three irregular regions once the accumulation process is finished.

Procedure 1 is decimation. A set of surviving nodes (i.e., parent nodes) is selected from the son level to build the parent level. This procedure is constrained by the following two rules (Meer, 1989): 1) Any two neighbor son nodes cannot both survive to the parent level and 2) any son node must have at least one parent node. Instead of the *random values* used in the stochastic pyramid decimation (SPD) algorithm (Jolion, 2003; Meer, 1989), this paper proposes a new recursive similarity-driven algorithm (i.e., the first extension), in which a son node will survive if it has the maximum *similarity* among its neighbors with the constraints of the aforementioned rules. The advantage is the improved segmentation performance since the nodes that can greatly represent their neighbors deterministically survive. As the second extension, *Bhattacharyya distance* (Bhattacharyya, 1943) is used to estimate the similarity between nodes at the same level (i.e., the strength of intra-level edges). One advantage is that the similarity measure is approximately scale-invariant during the accumulation process since Bhattacharyya distance takes into account the correlations of the data. The other advantage is that the probabilistic measure can improve the robustness to noise.

In *procedure 2*, the strength of inter-level edges is estimated. Each son node and its parent nodes are linked by inter-level edges. The strength of these edges is estimated in proportion to the corresponding intra-level strength at the son level by using the method in (Sharon et al., 2000).

Procedure 3 aims to estimate the aggregate features and covariances of each parent node based on the strength of inter-level edges by using the method in (Sharon et al., 2000).

The purpose of *procedure 4* is to search for neighbors of each parent node and simultaneously estimate the strength of intra-level edges at the parent level. As the third extension, a new

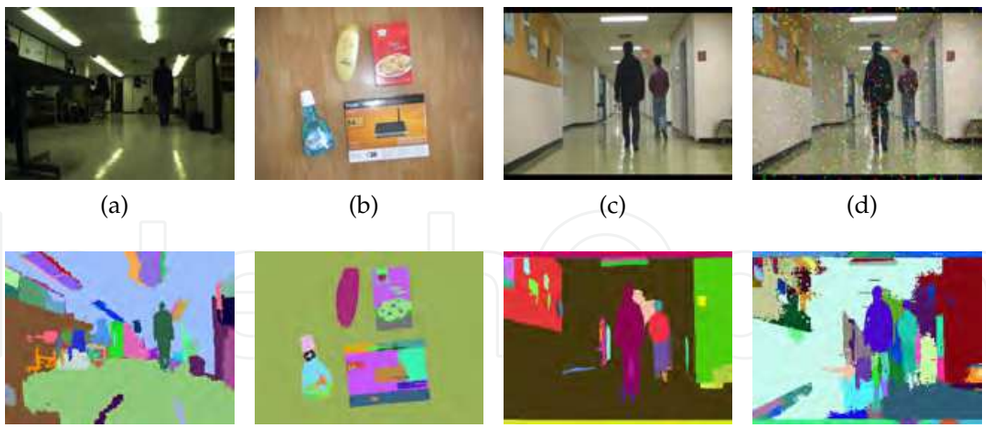


Fig. 4. Results of pre-attentive segmentation. (a)-(d) Original images, where (d) includes salt and pepper noise (noise density:0.1, patch size: 5×5 pixels). (e)-(h) Segmentation results. Each color represents one proto-object in these results.

neighbor search method is proposed by considering not only the graphic constraints but also the similarity constraints. A candidate node is selected as a neighbor of a center node if the similarity between them is above a predefined threshold. Since the similarity measure is scale-invariant, a fixed value of the threshold can be used for most pyramidal levels. The advantage of this method is the improved segmentation performance since the connections between nodes that are located at places with great transition are deterministically cut.
In the case that no neighbors are found for a node, it is labeled as a new proto-object. The construction of the full pyramid is finished once all nodes at a level have no neighbors. The membership of each node at the base level to each proto-object is iteratively calculated from the top pyramidal level to the base level. According to the membership, each node at the base level is finally labeled. The results of the pre-attentive segmentation are shown in Fig. 4.

5. Attentional selection

5.1 Bottom-up attention

The proposed bottom-up attention module is developed by extending Itti's model (Itti et al., 1998). Center-surround differences in terms of pre-attentive features are first calculated to simulate the competition in the spatial context:

$$\mathbf{F}'_f(s_c, s_s) = |\mathbf{F}_f(s_c) \ominus \mathbf{F}_f(s_s)| \tag{1}$$

where $\ominus$ denotes across-scale subtraction, consisting of interpolation of each feature at the surround scale to the center scale and point-by-point difference, $s_c = \{2,3,4\}$ and $s_s = s_c + \delta$ with $\delta = \{3,4\}$ represent the center scales and surround scales respectively, $f \in \{int, rg, by, o_\theta, ct\}$ with $\theta \in \{0°, 45°, 90°, 135°\}$, and $F'_f(s_c, s_s)$ denotes a center-surround difference map.
These center-surround differences in terms of the same feature dimension are then normalized and combined at scale 2, termed as *working scale* and denoted as s_{wk}, using across-scale

addition to yield a location-based conspicuity map of that feature dimension:

$$\mathbf{F}_f^s = \mathcal{N}\left(\frac{1}{6}\bigoplus_{s_c=2}^{4}\bigoplus_{s_s=s_c+3}^{s_c+4}\mathcal{N}\left(\mathbf{F}'_f(s_c,s_s)\right)\right) \tag{2}$$

where $\mathcal{N}$ is the normalization operator, $\bigoplus$ is across-scale addition, consisting of interpolation of each normalized center-surround difference to the working scale and point-by-point addition, $f \in \{int, rg, by, o_\theta, ct\}$, and $\mathbf{F}_f^s$ denotes a location-based conspicuity map.
All conspicuity maps are point-by-point added together to yield a location-based bottom-up saliency map $\mathbf{S}_{bu}$:

$$\mathbf{S}_{bu} = \mathcal{N}\left(\mathbf{F}_{ct}^s + \mathbf{F}_{int}^s + \frac{1}{2}(\mathbf{F}_{rg}^s + \mathbf{F}_{by}^s) + \frac{1}{4}\sum_\theta \mathbf{F}_{o_\theta}^s\right) \tag{3}$$

Given the following two assumptions: 1) the selection process guided by the space-based bottom-up attention is a random event, and 2) the sample space of this random event is composed of all spatial locations in the image, the salience of a spatial location can be used to represent the degree of belief that bottom-up attention selects that location. Therefore, the probability of a spatial location $\mathbf{r}_i$ being attended by the bottom-up attention mechanism can be estimated as:

$$p_{bu}(\mathbf{r}_i) = \frac{S_{bu}(\mathbf{r}_i)}{\sum_{\mathbf{r}_{i'}} S_{bu}(\mathbf{r}_{i'})}, \tag{4}$$

where $p_{bu}(\mathbf{r}_i)$ denotes the probability of a spatial location $\mathbf{r}_i$ being attended by the bottom-up attention, and the denominator $\sum_{\mathbf{r}_{i'}} S_{bu}(\mathbf{r}_{i'})$ is the normalizing constant.

5.2 Top-down attention

5.2.1 LTM task representations and task-relevant objects

The *task-relevant object* can be defined as an object whose occurrence is expected by the task. Consistent with the autonomous mental development (AMD) paradigm (Weng et al., 2001), this paper proposes that actions include external actions that operate effectors and internal actions that predict the next possible attentional state (i.e., attentional prediction). Since the proposed perception system is object-based, the attentional prediction can be seen as the task-relevant object. Thus this paper models the *LTM task representation* as the association between attentional states and attentional prediction and uses it to deduce the task-relevant object.
It can be further proposed that the LTM task representation can be modeled by using a first-order discrete Markov process (FDMP). The FDMP can be expressed as $p(a_{t+1}|a_t)$, where a_t denotes the attentional state at time t and a_{t+1} denotes the attentional prediction for time $t+1$. This definition means that the probability of each attentional prediction for the next time can be estimated given the attentional state at the current time. The discrete attentional states is composed of LTM object representations.

5.2.2 Task-relevant feature

According to the IC hypothesis, it is required to deduce the task-relevant feature from the task-relevant object. This paper defines the *task-relevant feature* as a property that can discriminate the object from others. Although several autonomous factors (e.g., rewards obtained from learning) could be used, this paper uses the *conspicuity* quantity since it

is one of the important intrinsic and innate properties of an object for measuring the discriminability. Through a training process that statistically encapsulates the conspicuity quantities obtained under different viewing conditions, a *salience descriptor* is achieved in the LTM object representation (See details in section 6.2 and section 6.3).
Therefore the salience descriptor is used to deduce the task-relevant feature by finding the feature dimension that has the greatest conspicuity. This deduction can be expressed as:

$$(f_{rel}, j_{rel}) = \arg \max_{f \in \{ct,int,rg,by,o_\theta\}} \max_{j \in \{1,2,\ldots,N_j\}} \frac{\overline{\mu}_f^{s,j}}{1 + \overline{\sigma}_f^{s,j}}, \tag{5}$$

where N_j denotes the number of parts when $f \in \{int, rg, by, o_\theta\}$ and $N_j = 1$ when $f = ct$, $\overline{\mu}_f^{s,j}$ and $\overline{\sigma}_f^{s,j}$ respectively denote the mean and STD of salience descriptors in terms of a feature f in the LTM representation of the task-relevant object, f_{rel} denotes the *task-relevant feature dimension*, and j_{rel} denotes the index of the *task-relevant part*. The LTM object representation can be seen in section 6.3.
In the proposed system, the most task-relevant feature is first selected for guiding top-down attention. If the post-attentive recognition shows that the attended object is not the target, then the next task-relevant feature is joined. This process does not stop until the attended object is verified or all features are used.

5.2.3 Attentional template

Given the task-relevant feature dimension, its *appearance descriptor* in the LTM representation of the task-relevant object is used to build an attentional template in WM so as to estimate top-down biases. The attentional template is denoted as $\mathbf{F}_f^t$, where $f \in \{ct, int, rg, by, o_\theta\}$. The appearance descriptor will be presented in section 6.3.

5.2.4 Estimation of top-down biases

Bayesian inference is used to estimate the location-based top-down bias, which represents the probability of a spatial location being an instance of the task-relevant object. It can be generally expressed as:

$$p_{td}(\mathbf{r}_i|\mathbf{F}_f^t) = \frac{p_{td}(\mathbf{F}_f^t|\mathbf{r}_i) \times p_{td}(\mathbf{r}_i)}{\sum_{\mathbf{r}_{i'}} p_{td}(\mathbf{F}_f^t|\mathbf{r}_{i'}) \times p_{td}(\mathbf{r}_{i'})}, \tag{6}$$

where $p_{td}(\mathbf{r}_i)$ denotes the prior probability of a location $\mathbf{r}_i$ being attended by the top-down attention, $p_{td}(\mathbf{F}_f^t|\mathbf{r}_i)$ denotes the observation likelihood, $p_{td}(\mathbf{r}_i|\mathbf{F}_f^t)$ is the posterior probability of the location $\mathbf{r}_i$ being attended by the top-down attention given the attentional template $\mathbf{F}_f^t$. Assuming that the prior probability $p_{td}(\mathbf{r}_i)$ is a uniform distribution, Eq. (6) can be simplified into estimating the observation likelihood $p_{td}(\mathbf{F}_f^t|\mathbf{r}_i)$. The detailed estimation of $p_{td}(\mathbf{F}_f^t|\mathbf{r}_i)$ for each feature dimension, including contour, intensity, red-green, blue-yellow and orientations can be seen in our previous object-based visual attention (OVA) model (Yu et al., 2010).

5.2.5 Discussion

Compared with existing top-down attention methods, e.g., (Navalpakkam & Itti, 2005; Rao & Ballard, 1995a), the proposed method has four advantages. The first advantage is effectiveness due to the use of both salience and appearance descriptors. These two descriptors reciprocally

aid each other: The salience descriptor guarantees that the task-relevant object can be effectively discriminated from distracters in terms of appearance, while the appearance descriptor can deal with the case that the task-relevant object and distracters have similar task-relevance values but different appearance values. The second advantage is efficiency. The computational complexity of (Rao & Ballard, 1995a) and our method can be approximated as $\mathcal{O}(d_h)$ and $\mathcal{O}(d_f^{few} d_l)$ respectively, where d_h denotes the dimension number of a high-level object representation, e.g., iconic representation (Rao & Ballard, 1995b) used in (Rao & Ballard, 1995a), d_l denotes the dimension number of a pre-attentive feature and d_f^{few} denotes the number of one or a few pre-attentive features used in our method. Since $d_h \gg d_f^{few} d_l$, the computation of our method is much cheaper. The third advantage is adaptability. As shown in (5), the task-relevant feature(s) can be autonomously deduced from the learned LTM representation such that the requirement of redesigning the representation of the task-relevant object for different tasks is eliminated. The fourth advantage is robustness. As shown in (6), the proposed method gives a bias toward the task-relevant object by using Bayes' rule, such that it is robust to work with noise, occlusion and a variety of viewpoints and illuminative effects.

5.3 Combination of bottom-up saliency and top-down biases

Assuming that bottom-up attention and top-down attention are two random events that are independent, the probability of an item being attended can be modeled as the probability of occurrence of either of these two events on that item. Thus, the probabilistic location-based attentional activation, denoted as $p_{attn}(\mathbf{r}_i)$, can be obtained by combining bottom-up saliency and top-down biases:

$$\begin{cases} p_{attn}(\mathbf{r}_i) = p_{bu}(\mathbf{r}_i) + p_{td}(\mathbf{r}_i|\{\mathbf{F}_f^t\}) - p_{bu}(\mathbf{r}_i) \times p_{td}(\mathbf{r}_i|\{\mathbf{F}_f^t\}) & \text{if } w_{bu} = 1 \text{ and } w_{td} = 1 \\ p_{attn}(\mathbf{r}_i) = p_{bu}(\mathbf{r}_i) & \text{if } w_{bu} = 1 \text{ and } w_{td} = 0 \\ p_{attn}(\mathbf{r}_i) = p_{td}(\mathbf{r}_i|\{\mathbf{F}_f^t\}) & \text{if } w_{bu} = 0 \text{ and } w_{td} = 1 \end{cases}, \quad (7)$$

where w_{bu} and w_{td} are two logic variables used as the conscious gating for bottom-up attention and top-down attention respectively and these two variables are set according to the task.

5.4 Proto-object based attentional activation

According to the IC hypothesis, it can be seen that a competitive advantage over an object is produced by directing attention to a spatial location in that object. Thus the probability of a proto-object being attended can be calculated using the *logic or* operator on the location-based probabilities. Furthermore, it can be assumed that two locations being attended are mutually exclusive according to the space-based attention theory (Posner et al., 1980). As a result, the probability of a proto-object $\mathbf{R}_g$ being attended, denoted as $p_{attn}(\mathbf{R}_g)$, can be calculated as:

$$p_{attn}(\mathbf{R}_g) = \frac{1}{N_g} \sum_{\mathbf{r}_i \in \mathbf{R}_g} p_{td}(\mathbf{r}_i|\mathbf{F}_f^t), \quad (8)$$

where $\mathbf{R}_g$ denotes a proto-object, N_g denotes the number of pixels in $\mathbf{R}_g$. The inclusion of $1/N_g$ is to eliminate the influence of the proto-object's size. The FOA is directed to the proto-object with maximal attentional activation.

6. Post-attentive perception

The flow chart of the post-attentive perception can be illustrated in Fig. 5. Four modules, as presented in section 3, are **interactive** during this stage.

6.1 Perceptual completion processing

This module works around the attended proto-object, denoted as $\mathbf{R}^1_{attn}$, to achieve the complete object region. It consists of two steps. The first step is recognition of the attended proto-object. This step explores LTM object representations in order to determine to which LTM object representation the attended proto-object belongs by using the post-attentive features. The extraction of post-attentive features and the recognition algorithm will be presented in section 6.2 and section 6.4 respectively. The matched LTM object representation, denoted as $\mathbf{O}_{attn}$, is then recalled from LTM.
The second step is completion processing:

1. If the local coding of $\mathbf{O}_{attn}$ includes multiple parts, several candidate proto-objects, which are spatially close to $\mathbf{R}^1_{attn}$, are selected from the current scene. They are termed as *neighbors* and denoted as a set $\{\mathbf{R}_n\}$.
2. The local post-attentive features are extracted in each $\mathbf{R}_n$.
3. Each $\mathbf{R}_n$ is recognized using the local post-attentive features and the matched LTM object representation $\mathbf{O}_{attn}$. If it is recognized as a part of $\mathbf{O}_{attn}$, it will be labeled as a part of the attended object. Otherwise, it will be eliminated.
4. Continue *item 2* and *item 3* iteratively until all neighbors have been checked.

These labeled proto-objects constitute the complete region of the attended object, which is denoted as a set $\{\mathbf{R}_{attn}\}$.

6.2 Extraction of post-attentive features

Post-attentive features $\tilde{\mathbf{F}}$ are estimated by using the statistics within the attended object. They consist of *global post-attentive features* $\tilde{\mathbf{F}}_{gb}$ and *local post-attentive features* $\tilde{\mathbf{F}}_{lc}$. Each $\tilde{\mathbf{F}}$ consists of *appearance component* $\tilde{\mathbf{F}}^a$ and *salience component* $\tilde{\mathbf{F}}^s$.

6.2.1 Local post-attentive features

Each proto-object, denoted as $\mathbf{R}^m_{attn}$, in the complete region being attended (i.e., $\mathbf{R}^m_{attn} \in \{\mathbf{R}_{attn}\}$) is the unit for estimating local post-attentive features. They can be estimated as a set that can be expressed as: $\{\tilde{\mathbf{F}}_{lc}\} = \{(\tilde{\mathbf{F}}^a_{lc}(\mathbf{R}^m_{attn}), \tilde{\mathbf{F}}^s_{lc}(\mathbf{R}^m_{attn}))^T\}_{\forall \mathbf{R}^m_{attn} \in \{\mathbf{R}_{attn}\}}$.
The appearance components in an entry $\tilde{\mathbf{F}}_{lc}$, denoted as $\tilde{\mathbf{F}}^a_{lc} = \{\tilde{\mathbf{F}}^a_f\}$ with $f \in \{int, rg, by, o_\theta\}$, are estimated by using the mean $\tilde{\mu}^{a,m}_f$ of $\mathbf{R}^m_{attn}$ in terms of f, i.e., $\tilde{\mathbf{F}}^a_f(\mathbf{R}^m_{attn}) = \tilde{\mu}^{a,m}_f$.
The salience components, denoted as $\tilde{\mathbf{F}}^s_{lc} = \{\tilde{\mathbf{F}}^s_f\}$ with $f \in \{int, rg, by, o_\theta\}$, can be estimated using the mean of conspicuity $\tilde{\mu}^{s,m}_f$ of a $\mathbf{R}^m_{attn}$ in terms of f, i.e., $\tilde{F}^s_f(\mathbf{R}^m_{attn}) = \tilde{\mu}^{s,m}_f$. The conspicuity quantity F^s_f in terms of f is calculated using (2).

6.2.2 Global post-attentive features

The global post-attentive feature $\tilde{\mathbf{F}}_{gb}$ is estimated after the complete region of the attended object, i.e., $\{\mathbf{R}_{attn}\}$, is obtained. Since the active contour technique (Blake & Isard, 1998; MacCormick, 2000) is used to represent a contour in this paper, the estimation of $\tilde{\mathbf{F}}_{gb}$ includes

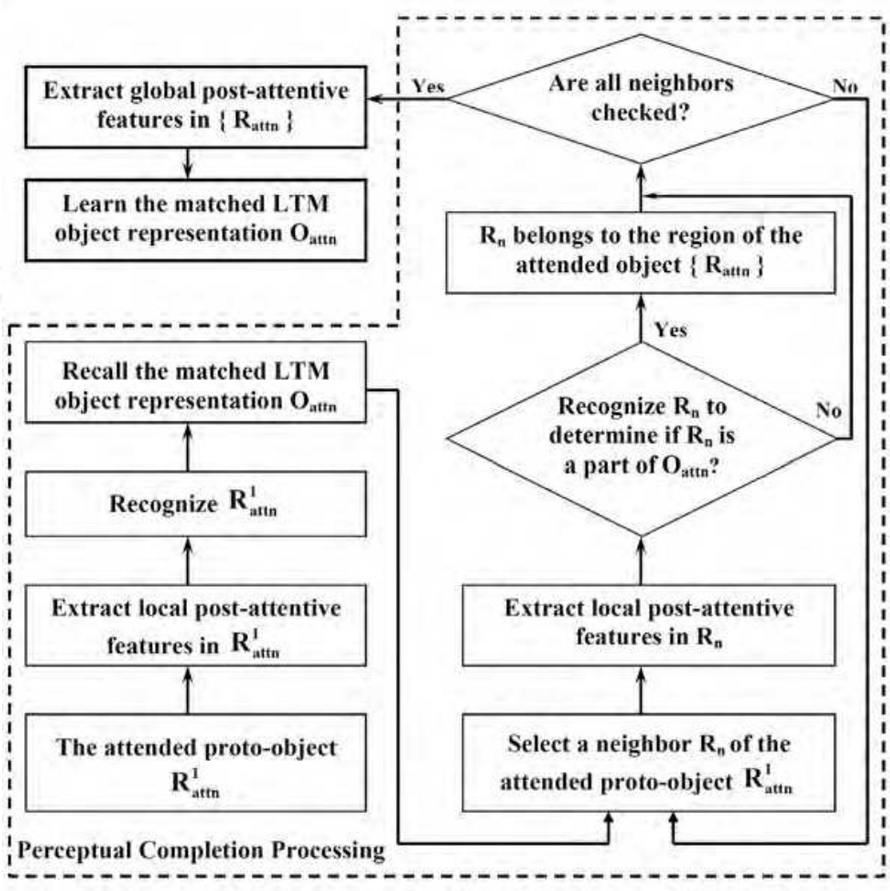


Fig. 5. The flowchart of the post-attentive perception stage.

two steps. The first step is to extract control points, denoted as a set $\{\mathbf{r}_{cp}\}$, of the attended object's contour by using the method in our previous work (Yu et al., 2010). That is, each control point is an entry in the set $\{\tilde{\mathbf{F}}_{gb}\}$. The second step is to estimate the appearance and salience components at these control points, i.e., $\{\tilde{\mathbf{F}}_{gb}\} = \{(\tilde{\mathbf{F}}^a_{gb}(\mathbf{r}_{cp}), \tilde{\mathbf{F}}^s_{gb}(\mathbf{r}_{cp}))^T\}_{\forall \mathbf{r}_{cp}}$. The appearance component of an entry consists of spatial coordinates in the reference frame at a control point, i.e., $\tilde{\mathbf{F}}^a_{gb}(\mathbf{r}_{cp}) = (x_{\mathbf{r}_{cp}} \; y_{\mathbf{r}_{cp}})^T$. The salience component of an entry is built by using the conspicuity value $F^s_{ct}(\mathbf{r}_{cp})$ in terms of pre-attentive contour feature at a control point, i.e., $\tilde{\mathbf{F}}^s_{gb}(\mathbf{r}_{cp}) = F^s_{ct}(\mathbf{r}_{cp})$.

6.3 Development of LTM object representations

The LTM object representation also consists of the local coding (denoted as $\mathbf{O}_{lc}$) and global coding (denoted as $\mathbf{O}_{gb}$). Each coding also consists of *appearance descriptors* (denoted as $\mathbf{O}^a$) and *salience descriptors* (denoted as $\mathbf{O}^s$). The PNN (Specht, 1990) is used to build them.

6.3.1 PNN of local coding

The PNN of a local coding $\mathbf{O}_{lc}$ (termed as a *local PNN*) includes three layers. The input layer receives the local post-attentive feature vector $\tilde{\mathbf{F}}_{lc}$. Each radial basis function (RBF) at the

hidden layer represents a part of the learned object and thereby this layer is called a *part layer*. The output layer is a probabilistic mixture of all parts belonging to the object and thereby this layer is called an *object layer*.
The probability distribution of a RBF at the part layer of the local PNN can be expressed as:

$$\begin{aligned} p_j^k(\tilde{\mathbf{F}}_{lc}) &= \mathcal{G}(\tilde{\mathbf{F}}_{lc}; \boldsymbol{\mu}_j^k, \boldsymbol{\Sigma}_j^k) \\ &= \frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}_j^k|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\tilde{\mathbf{F}}_{lc} - \boldsymbol{\mu}_j^k)^T (\boldsymbol{\Sigma}_j^k)^{-1} (\tilde{\mathbf{F}}_{lc} - \boldsymbol{\mu}_j^k)\} \end{aligned} \tag{9}$$

where $\mathcal{G}$ denotes the Gaussian distribution, $\boldsymbol{\mu}_j^k$ and $\boldsymbol{\Sigma}_j^k$ denote the mean vector and covariance matrix of a RBF, j is the index of a part, k is the index of an object in LTM, and d is the dimension number of a local post-attentive feature $\tilde{\mathbf{F}}_{lc}$. Since all feature dimensions are assumed to be independent, $\boldsymbol{\Sigma}_j^k$ is a diagonal matrix and standard deviation (STD) values of all feature dimensions of a RBF can constitute an STD vector $\boldsymbol{\sigma}_j^k$.
The probabilistic mixture estimation $r^k(\tilde{\mathbf{F}}_{lc})$ at the object layer can be expressed as:

$$r^k(\tilde{\mathbf{F}}_{lc}) = \sum_i \pi_j^k p_j^k(\tilde{\mathbf{F}}_{lc}), \tag{10}$$

where π_j^k denotes the contribution of part j to object k, which holds $\sum_j \pi_j^k = 1$.

6.3.2 PNN of global coding

The PNN for a global coding $\mathbf{O}_{gb}$ (termed as a *global PNN*) also includes three layers. The input layer receives the global post-attentive feature vector $\tilde{\mathbf{F}}_{gb}$. Each node of the hidden layer is a control point along the contour and thereby this layer is called a *control point layer*. The output layer is a probabilistic combination of all control points belonging to the object and thereby this layer is called an *object layer*. The mathematical expression of the global PNN is similar to the local PNN.

6.3.3 Learning of LTM object representations

Since the number of nodes (i.e., the numbers of parts and control points) is unknown and might be dynamically changed during the training course, this paper proposes a dynamical learning algorithm by using both the maximum likelihood estimation (MLE) and a Bayes' classifier to update the local and global PNNs at each time. This proposed dynamical learning algorithm can be summarized as follows. The Bayes' classifier is used to classify the training pattern to an existing LTM pattern. If the training pattern can be classified to an existing LTM pattern at the part level in a local PNN or at the control point level in a global PNN, both appearance and salience descriptors of this existing LTM pattern are updated using MLE. Otherwise, a new LTM pattern is created. Two thresholds τ_1 and τ_2 are introduced to determine the minimum correct classification probability to an existing part and an existing control point respectively. Algorithm 1 shows the learning routine of global and local codings. In the algorithm, a_j^k denotes the occurrence number of an existing pattern indexed by j of object k and it is initialized by 0, N_k denotes the number of parts in the local PNN or control points in the global PNN of object k, $.^2$ denotes the element-by-element square operator, and σ_{init} is a predefined STD value when a new pattern is created.

Algorithm 1 Learning Routine of Local and Global Codings

Given a local or global training pattern $(\tilde{\mathbf{F}}_{lc}, k)$ or $(\tilde{\mathbf{F}}_{gb}, k)$:
Set $\tilde{\mathbf{F}} = \tilde{\mathbf{F}}_{lc}$ or $\tilde{\mathbf{F}} = \tilde{\mathbf{F}}_{gb}$;
Recognize $\tilde{\mathbf{F}}$ to obtain a recognition probability $p_i^k(\tilde{\mathbf{F}})$;
if $p_j^k(\tilde{\mathbf{F}}) \geq \tau_1$ or $\geq \tau_2$ **then**
    // Update part j of object k
    $\sigma_{temp} = \left[a_j^k(\sigma_j^k).^2 + a_j^k(\mu_j^k).^2 + (\tilde{\mathbf{F}}).^2\right] / (a_j^k + 1)$; *// Prepare for updating the STD*
    $\boldsymbol{\mu}_j^k = (a_j^k \boldsymbol{\mu}_j^k + \tilde{\mathbf{F}}) / (a_j^k + 1)$; *// Update the mean vector*
    $\sigma_j^k = [\sigma_{temp}^d - (\mu_j^k).^2].^{-\frac{1}{2}}$; *// Update the STD*
    $a_j^k = a_j^k + 1$; *// Increment the occurrence number*
else
    // Create a new part i of object k
    Set $N_k = N_k + 1$; $i = N_k$;
    $\boldsymbol{\mu}_j^k = \tilde{\mathbf{F}}$; $\boldsymbol{\sigma}_j^k = \boldsymbol{\sigma}_{init}$; $a_j^k = 1$; *// Set the initial mean, STD and occurrence number*
end if
$\forall j$: $\pi_j^k = a_j^k / \sum_{j'} a_{j'}^k$. *// Normalize weights π*

6.4 Object recognition

Due to the page limitation, the object recognition module can be summarized as follows. It can be modeled at two levels. The first one is the object level. The purpose of this level is to recognize to which LTM object an attended pattern belongs. The second one is the part level or control point level. Recognition at this level is performed given an LTM object to which the attended pattern belongs. Thus, the purpose of this level is to recognize to which part in a local PNN or to which control point in a global PNN an attended pattern belongs. At each level, object recognition can generally be modeled as a decision unit by using Bayes' theorem. Assuming that the prior probability is equal for all LTM patterns at each level, the observation likelihood can be seen as the posterior probability.

7. Experiments

This proposed autonomous visual perception system is tested in the task of object detection. The unconscious perception path (i.e., the bottom-up attention module) can be used to detect a salient object, such as a landmark, whereas the conscious perception path (i.e., the top-down attention module) can be used to detect the task-relevant object, i.e., the expected target. Thus the unconscious and conscious aspects are tested in two robotics tasks respectively: One is detecting a salient object and the other is detecting a task-relevant object.

7.1 Detecting a salient object

The salient object is an unusual or unexpected object and the current task has no prediction about its occurrence. There are three objectives in this task. The first objective is to illustrate the unconscious capability of the proposed perception system. The second objective is to show the advantages of using object-based visual attention for perception by comparing it with the space-based visual attention methods. The third objective is to show the advantage of integrating the contour feature into the bottom-up competition module. The result is that an object that has a conspicuous shape compared with its neighbors can be detected. Two

experiments are shown in this section, including the detection of an object that is conspicuous in colors and in contour respectively.

7.1.1 Experimental setup

Artificial images are used in the experiments. The frame size of all images is 640×480 pixels. In order to show the robustness of the proposed perception system, these images are obtained using different settings, including noise, spatial transformation and changes of lighting. The noisy images are manually obtained by adding salt and pepper noise patches (noise density: $0.1 \sim 0.15$, patch size: 10×10 pixels $\sim 15 \times 15$ pixels) into original r, g and b color channels respectively. The experimental results are compared with the results of Itti's model (i.e., space-based bottom-up attention) (Itti et al., 1998) and Sun's model (i.e., object-based bottom-up attention) (Sun & Fisher, 2003).

7.1.2 An object conspicuous in colors

The first experiment is detecting an object that is conspicuous to its neighbors in terms of colors and all other features are approximately the same between the object and its neighbors. The experimental results are shown in Fig. 6. The salient object is the red ball in this experiment. Results of the proposed perception system are shown in Fig. 6(d), which indicate that this proposed perception system can detect the object that is conspicuous to its neighbors in terms of colors in different settings. Results of Itti's model and Sun's model are shown in Fig. 6(e) and Fig. 6(f) respectively. It can be seen that Itti's model fails to detect the salient object when noise is added to the image, as shown in column 2 in Fig. 6(e). This indicates that the proposed object-based visual perception system is more robust to noise than the space-based visual perception methods.

7.1.3 An object conspicuous in contour

The second experiment is detecting an object that is conspicuous to its neighbors in terms of contour and all other features are approximately the same between the object and its neighbors. The experimental results are shown in Fig. 7. In this experiment, the salient object is the triangle. Detection results of the proposed perception system are shown in Fig. 7(d), which indicate that the proposed perception system can detect the object that is conspicuous to its neighbors in terms of contour in different settings. Detection results of Itti's model and Sun's model are shown in Fig. 7(e) and Fig. 7(f) respectively. It can be seen that both Itti's model and Sun's model fail to detect the salient object when noise is added to the image, as shown in column 2 in Fig. 7(e) and Fig. 7(f) respectively. This experiment indicates that the proposed object-based visual perception system is capable of detecting the object conspicuous in terms of contour in different settings due to the inclusion of contour conspicuity in the proposed bottom-up attention module.

7.2 Detecting a task-relevant object

It is an important ability for robots to accurately detect a task-relevant object (i.e., target) in the cluttered environment. According to the proposed perception system, the detection procedure consists of two phases: a learning phase and a detection phase. The objective of the learning phase is to develop the LTM representation of the target. The objective of the detection phase is to detect the target by using the learned LTM representation of the target. The detection phase can be implemented as a two-stage process. The first stage is attentional selection: The

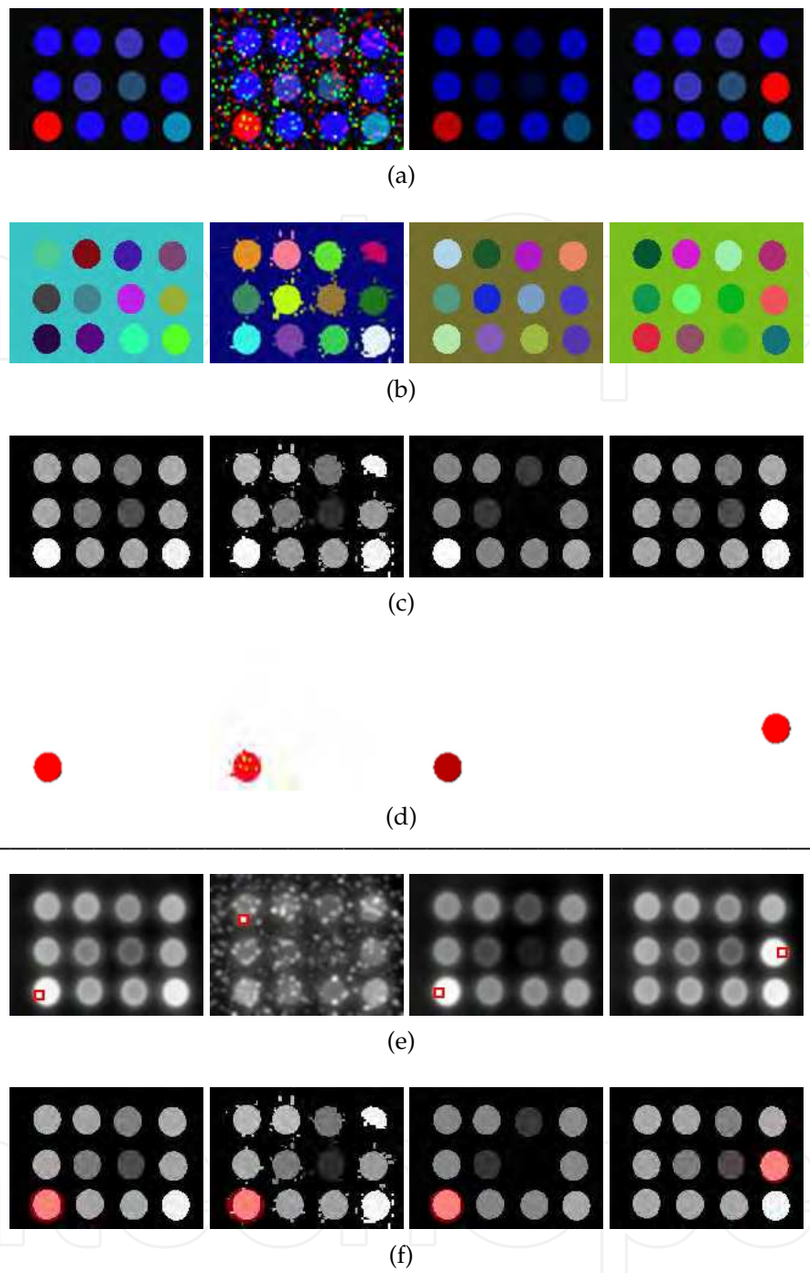


Fig. 6. Detection of a salient object, which is conspicuous to its neighbors in terms of colors. Each column represents a type of experimental setting. Column 1 is a typical setting. Column 2 is a noise setting of column 1. Column 3 is a different lighting setting with respect to column 1. Column 4 is a spatial transformation setting with respect to column 1. Row (a): Original input images. Row (b): Pre-attentive segmentation. Each color represents one proto-object. Row (c): Proto-object based attentional activation map. Row (d): The complete region being attended. Row (e): Detection results using Itti's model. The red rectangles highlight the attended location. Row (f): Detection results using Sun's model. The red circles highlight the attended object.

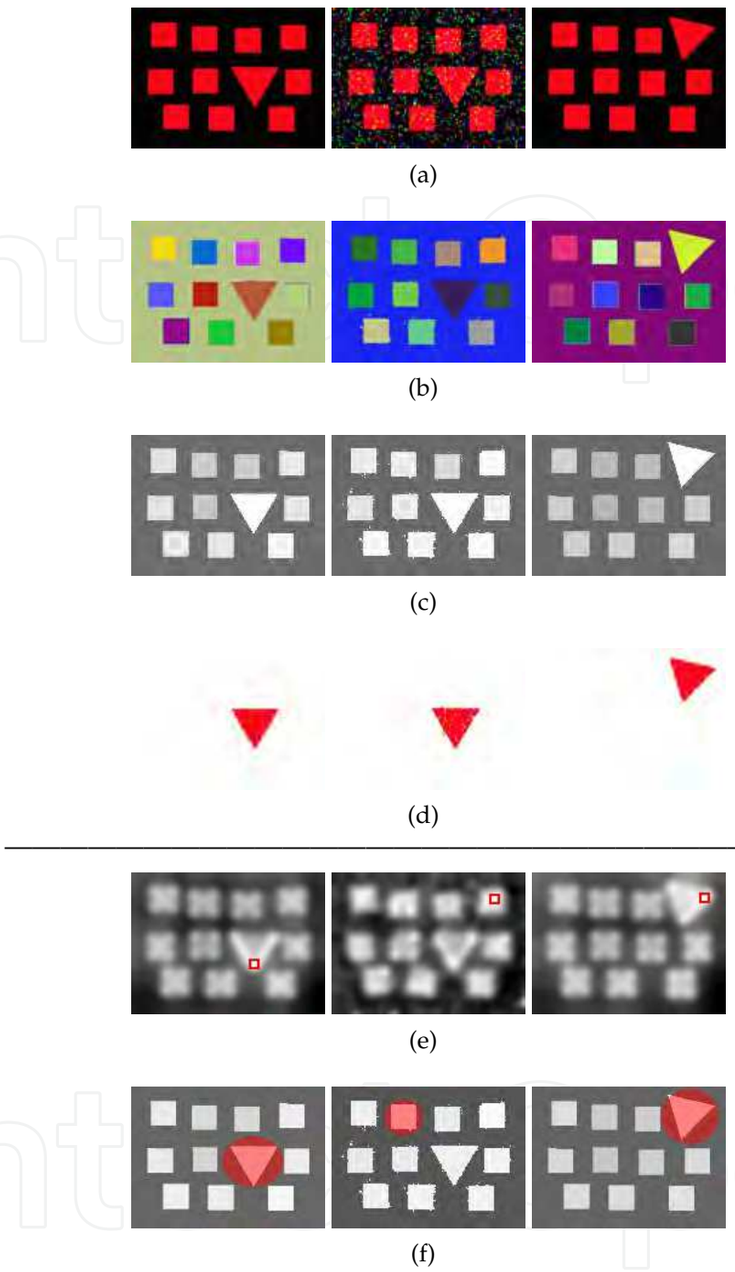


Fig. 7. Detection of a salient object, which is conspicuous to its neighbors in terms of contour. Each column represents a type of experimental setting. Column 1 is a typical setting. Column 2 is a noise setting of column 1. Column 3 is a spatial transformation setting with respect to column 1. Row (a): Original input images. Row (b): Pre-attentive segmentation. Each color represents one proto-object. Row (c): Proto-object based attentional activation map. Row (d): The complete region being attended. Row (e): Detection results using Itti's model. The red rectangles highlight the attended location. Row (f): Detection results using Sun's model. The red circles highlight the attended object.

task-relevant feature(s) of the target is used to guide attentional selection through top-down biasing to obtain an attended object. The second stage is post-attentive recognition: The attended object is recognized using the target's LTM representation to check if it is the target. If not, another procedure of attentional selection is performed by using more task-relevant features.

7.2.1 Experimental setup

Two objects are used to test the proposed method of detecting a task-relevant object: a book and a human. Images and videos are obtained under different settings, including noise, transformation, lighting changes and occlusion. For training for the book, 20 images are used. For testing for the book, 50 images are used. The size of each image is 640×480 pixels. For detecting the human, three videos are obtained by a moving robot. Two different office environments have been used. Video 1 and video 2 are obtained in office scene 1 with low and high lighting conditions respectively. Video 3 is obtained in office scene 2. All three videos contain a total of 650 image frames, in which 20 image frames are selected from video 1 and video 2 for training and the rest of the 630 image frames are used for testing. The size of each frame in these videos is 1024×768 pixels. It is important to note that each test image includes not only a target but also various distracters. The noisy images are manually obtained by adding salt and pepper noise patches (noise density: 0.1, patch size: 5×5 pixels) into original r, g and b color channels respectively.
The results of the proposed method are compared with the results of Itti's model (Itti et al., 1998) (i.e., a space-based bottom-up attention model) and Navalpakkam's model (Navalpakkam & Itti, 2005) (i.e., a space-based top-down attention model) respectively.

7.2.2 Task 1

The first task is to detect the book that has multiple parts. The learned LTM representation of the book is shown in Table 1, which has shown that the book has two parts and the blue-yellow feature in the first part can be deduced as the task-relevant feature dimension since the value $\mu^s/(1+\sigma^s)$ of this feature is maximal. Detection results of the proposed perception system are shown in Fig. 8(d). It can be seen that the book is successfully detected. Results of Itti's model and Navalpakkam's model, as shown in Fig. 8(e) and Fig. 8(f) respectively, show that these models fail to detect the target in some cases.

7.2.3 Task 2

The second task is to detect a human. Table 2 has shown that the human has two parts (including face and body) and the contour feature can be deduced as the task-relevant feature dimension since the value $\mu^s/(1+\sigma^s)$ of this feature is maximal. Detection results of the proposed perception system are shown in Fig. 9(d). It can be seen that the human is successfully detected. Results of Itti's model and Navalpakkam's model, as shown in Fig. 9(e) and Fig. 9(f) respectively, show that these models fail to detect the target in most cases.

7.2.4 Performance evaluation

Performance of detecting task-relevant objects is evaluated using true positive rate (TPR) and false positive rate (FPR), which are calculated as:

$$TPR = TP/nP, \tag{11}$$

f	j	μ^a	σ^a	μ^s	σ^s	$\mu^s/(1+\sigma^s)$
ct	1	-		75.0	19.7	3.6
int	1	106.6	5.8	27.9	14.5	1.8
rg	1	22.1	8.7	199.6	18.2	10.4
by	1	-108.0	9.1	215.6	8.7	**22.2**
o_{0°	1	N/A	N/A	41.8	9.8	3.9
o_{45°	1	N/A	N/A	41.4	12.8	3.0
o_{90°	1	N/A	N/A	34.7	16.3	2.0
o_{135°	1	N/A	N/A	46.5	15.7	2.8
int	2	60.5	8.2	80.0	5.7	11.9
rg	2	0.4	4.3	18.3	6.4	2.5
by	2	120.8	6.7	194.7	8.1	21.4
o_{0°	2	N/A	N/A	48.5	11.1	4.0
o_{45°	2	N/A	N/A	53.8	9.9	4.9
o_{90°	2	N/A	N/A	38.4	14.6	2.5
o_{135°	2	N/A	N/A	59.4	20.3	2.8

Table 1. Learned LTM object representation of the book. f denotes a pre-attentive feature dimension. j denotes the index of a part. The definitions of μ^a, σ^a, μ^s and σ^s can be seen in section 5.2.2.

f	j	μ^a	σ^a	μ^s	σ^s	$\mu^s/(1+\sigma^s)$
ct	1	-		68.3	6.9	**8.6**
int	1	28.4	21.7	18.8	13.9	1.3
rg	1	-7.0	7.1	28.6	10.8	2.4
by	1	10.9	5.4	48.4	10.9	4.1
o_{0°	1	N/A	N/A	33.4	6.7	4.3
o_{45°	1	N/A	N/A	39.8	11.4	3.2
o_{90°	1	N/A	N/A	37.4	6.1	5.3
o_{135°	1	N/A	N/A	37.5	13.5	2.6
int	2	52.0	12.5	25.6	15.6	1.5
rg	2	-2.3	17.4	49.5	18.8	2.5
by	2	-29.3	6.9	60.4	22.3	2.6
o_{0°	2	N/A	N/A	12.1	6.6	1.6
o_{45°	2	N/A	N/A	16.5	8.3	1.8
o_{90°	2	N/A	N/A	15.0	7.9	1.7
o_{135°	2	N/A	N/A	17.2	8.1	1.9

Table 2. Learned LTM object representation of the human. f denotes a pre-attentive feature dimension. j denotes the index of a part. The definitions of μ^a, σ^a, μ^s and σ^s can be seen in section 5.2.2.

$$FPR = FP/nN, \tag{12}$$

where nP and nN are numbers of positive and negative objects respectively in the testing image set, TP and FP are numbers of true positives and false positives. The positive object is the target to be detected and the negative objects are distracters in the scene.
Detection performance of the proposed perception system and other visual attention based methods is shown in Table 3. Note that “Naval’s" represents Navalpakkam’s method.

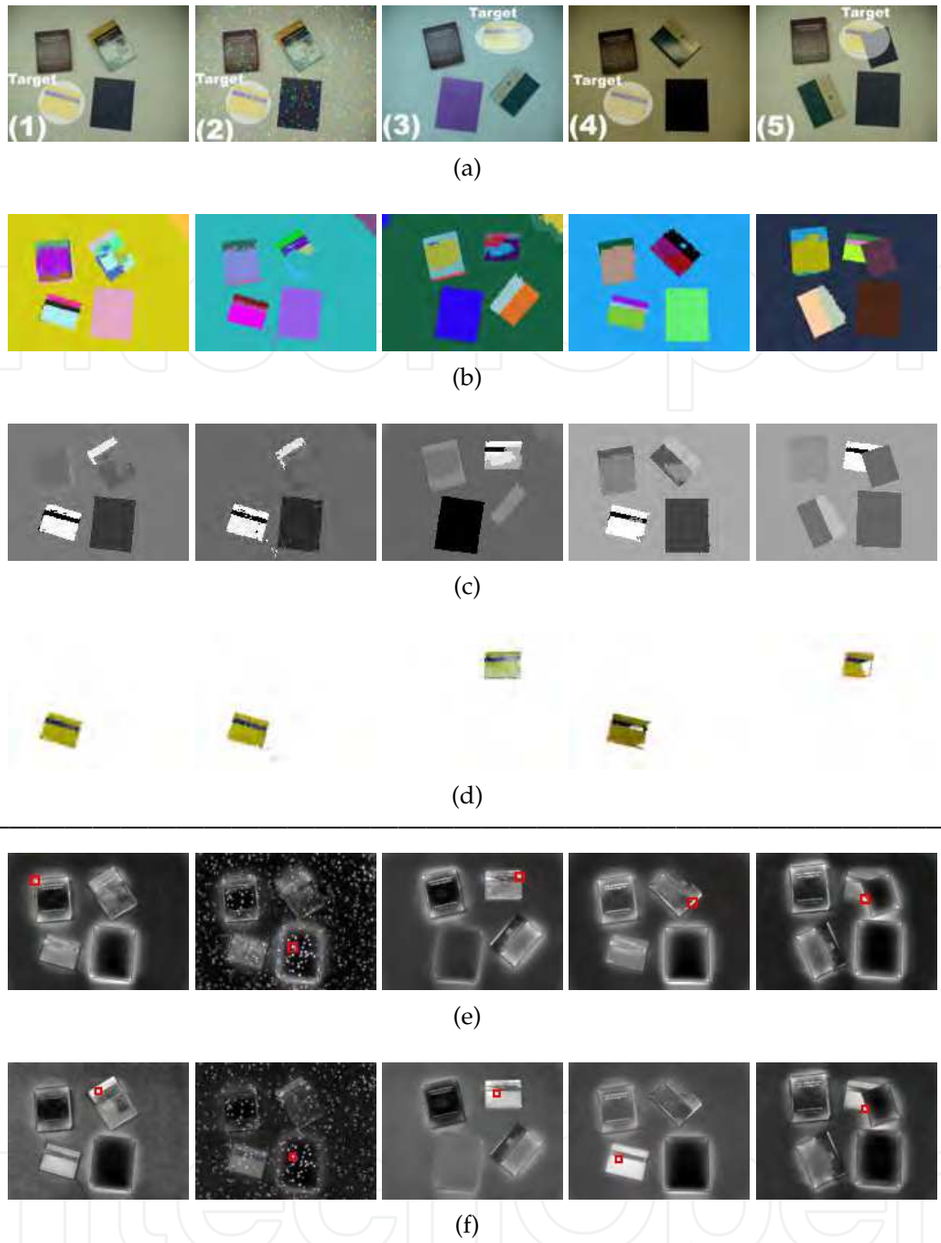


Fig. 8. Detection of the book. Each column represents a type of experimental setting. Column 1 is a typical setting. Column 2 is a noise setting of column 1. Column 3 is a spatial transformation (including translation and rotation) setting with respect to column 1. Column 4 is a different lighting setting with respect to column 1. Column 5 is an occlusion setting. Row (a): Original input images. Row (b): Pre-attentive segmentation. Each color represents one proto-object. Row (c): Proto-object based attentional activation map. Brightness represents the attentional activation value. Row (d): The complete region of the target. The red contour in the occlusion case represents the illusory contour (Lee & Nguyen, 2001), which shows the post-attentive perceptual completion effect. Row (e): Detection results using Itti's model. The red rectangle highlights the most salient location. Row (f): Detection results using Navalpakkam's model. The red rectangle highlights the most salient location.

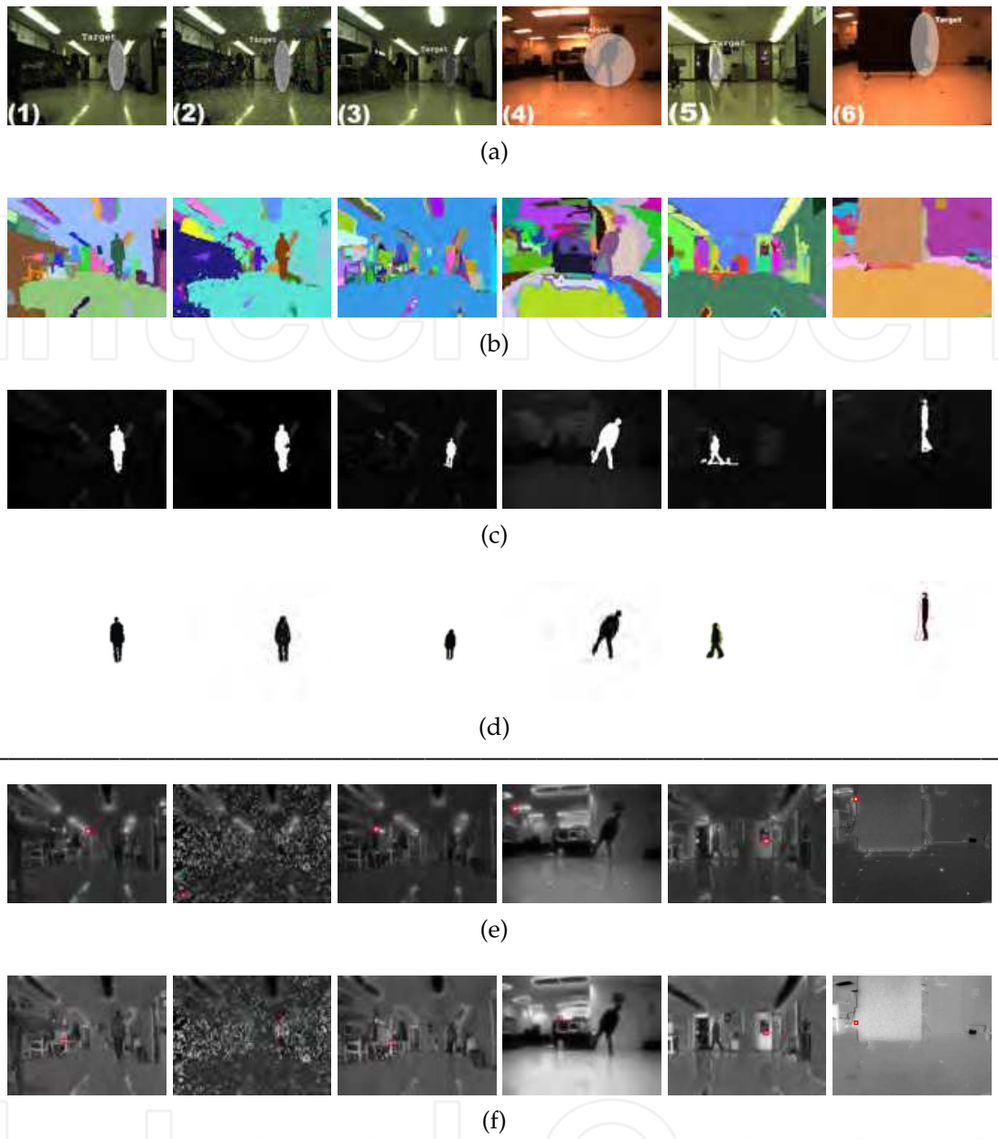


Fig. 9. Detection of the human in the cluttered environment. Each column represents a type of experimental setting. Column 1 is a typical setting (from video 1). Column 2 is a noise setting of column 1. Column 3 is a scaling setting with respect to column 1 (from video 1). Column 4 is a rotation setting with respect to column 1 (from video 3). Column 5 is a different lighting setting with respect to column 1 (from video 2). Column 6 is an occlusion setting (from video 3). Row (a): Original input images. Row (b): Pre-attentive segmentation. Each color represents one proto-object. Row (c): Proto-object based attentional activation map. Brightness represents the attentional activation value. Row (d): The complete region of the target. The red contour in the occlusion case represents the illusory contour (Lee & Nguyen, 2001), which shows the post-attentive perceptual completion effect. Row (e): Detection results using Itti's model. The red rectangle highlights the most salient location. Row (f): Detection results using Navalpakkam's model. The red rectangle highlights the most salient location.

Task	Method	*TP*	*FP*	*nP*	*nN*	*TPR* (%)	*FPR* (%)
1	Proposed	47	3	50	244	94.00	1.23
	Itti's	16	34	50	244	32.00	13.93
	Naval's	41	9	50	244	82.00	3.69
2	Proposed	581	49	630	30949	92.22	0.16
	Itti's	5	625	630	30949	0.79	2.02
	Naval's	36	594	630	30949	5.71	1.92

Table 3. Performance of detecting task-relevant objects.

8. Conclusion

This paper has presented an autonomous visual perception system for robots using the object-based visual attention mechanism. This perception system provides the following four contributions. The first contribution is that the attentional selection stage supplies robots with the cognitive capability of knowing how to perceive the environment according to the current task and situation, such that this perception system is adaptive and general to any task and environment. The second contribution is the top-down attention method using the IC hypothesis. Since the task-relevant feature(s) are conspicuous, low-level and statistical, this top-down biasing method is more effective, efficient and robust than other methods. The third contribution is the PNN based LTM object representation. This LTM object representation can probabilistically embody various instances of that object, such that it is robust and discriminative for top-down attention and object recognition. The fourth contribution is the pre-attentive segmentation algorithm. This algorithm extends the irregular pyramid techniques by integrating a scale-invariant probabilistic similarity measure, a similarity-driven decimation method and a similarity-driven neighbor search method. It provides rapid and satisfactory results of pre-attentive segmentation for object-based visual attention. Based on these contributions, this perception system has been successfully tested in the robotic task of object detection under different experimental settings.
The future work includes the integration of the bottom-up attention in the temporal context and experiments of the combination of bottom-up and top-down attention.

9. References

Aziz, M. Z., Mertsching, B., Shafik, M. S. E.-N. & Stemmer, R. (2006). Evaluation of visual attention models for robots, *Proceedings of the 4th IEEE Conference on Computer Vision Systems*, p. 20.

Backer, G., Mertsching, B. & Bollmann, M. (2001). Data- and model-driven gaze control for an active-vision system, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23(12): 1415–1429.

Baluja, S. & Pomerleau, D. (1997). Dynamic relevance: Vision-based focus of attention using artificial neural networks, *Artificial Intelligence* 97: 381–395.

Belardinelli, A. & Pirri, F. (2006). A biologically plausible robot attention model, based on space and time, *Cognitive Processing* 7(Supplement 5): 11–14.

Belardinelli, A., Pirri, F. & Carbone, A. (2006). Robot task-driven attention, *Proceedings of the International Symposium on Practical Cognitive Agents and Robots*, pp. 117–128.

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions, *Bulletin of the Calcutta Mathematical Society* 35: 99–109.

Blake, A. & Isard, M. (1998). *Active Contours: The Application of Techniques from Graphics, Vision, Control Theory and Statistics to Visual Tracking of Shapes in Motion*, Springer-Verlag New York, Secaucus, NJ.

Breazeal, C., Edsinger, A., Fitzpatrick, P. & Scassellati, B. (2001). Active vision for sociable robots, *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 31(5): 443–453.

Burt, P. J. & Adelson, E. H. (1983). The laplacian pyramid as a compact image code, *IEEE Transactions on Communications* 31(4): 532–540.

Carbone, A., Finzi, A. & Orlandini, A. (2008). Model-based control architecture for attentive robots in rescue scenarios, *Autonomous Robots* 24(1): 87–120.

Carpenter, G. A. & Grossberg, S. (2003). Adaptive resonance theory, *in* M. A. Arbib (ed.), *The Handbook of Brain Theory and Neural Networks, Second Edition*, MIT Press, Cambridge, MA, pp. 87–90.

Chikkerur, S., Serre, T., Tan, C. & Poggio, T. (2010). What and where: A bayesian inference theory of attention, *Visual Research* 50(22): 2233–2247.

Desimone, R. & Duncan, J. (1995). Neural mechanisms of selective visual attention, *Annual Reviews of Neuroscience* 18: 193–222.

Duncan, J. (1984). Selective attention and the organization of visual information, *Journal of Experimental Psychology: General* 113(4): 501–517.

Duncan, J. (1998). Converging levels of analysis in the cognitive neuroscience of visual attention, *Philosophical Transactions of The Royal Society Lond B: Biological Sciences* 353(1373): 1307–1317.

Duncan, J., Humphreys, G. & Ward, R. (1997). Competitive brain activity in visual attention, *Current Opinion in Neurobiology* 7(2): 255–261.

Frintrop, S. (2005). *VOCUS: A visual attention system for object detection and goal-directed search*, PhD thesis, University of Bonn, Germany.

Frintrop, S. & Jensfelt, P. (2008). Attentional landmarks and active gaze control for visual slam, *IEEE Transactions on Robotics* 24(5): 1054–1065.

Frintrop, S. & Kessel, M. (2009). Most salient region tracking, *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1869–1874.

Greenspan, A. G., Belongie, S., Goodman, R., Perona, P., Rakshit, S. & Anderson, C. H. (1994). Overcomplete steerable pyramid filters and rotation invariance, *Proceedings of IEEE Computer Vision and Pattern Recognition*, pp. 222–228.

Grossberg, S. (2005). Linking attention to learning, expectation, competition, and consciousness, *in* L. Itti, G. Rees & J. Tsotsos (eds), *Neurobiology of Attention*, Elsevier, San Diego, CA, pp. 652–662.

Grossberg, S. (2007). Consciousness clears the mind, *Neural Networks* 20: 1040–1053.

Hoya, T. (2004). Notions of intuition and attention modeled by a hierarchically arranged generalized regression neural network, *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics* 34(1): 200–209.

Itti, L. & Baldi, P. (2009). Bayesian surprise attracts human attention, *Visual Research* 49(10): 1295–1306.

Itti, L., Koch, C. & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(11): 1254–1259.

Jolion, J. M. (2003). Stochastic pyramid revisited, *Pattern Recognition Letters* 24(8): 1035–1042.

Lee, T. S. & Nguyen, M. (2001). Dynamics of subjective contour formation in the early visual cortex, *Proceedings of the Natual Academy of Sciences of the United States of America* 98(4): 1907–1911.

MacCormick, J. (2000). *Probabilistic modelling and stochastic algorithms for visual localisation and tracking*, PhD thesis, Department of Engineering Science, University of Oxford.

Maier, W. & Steinbach, E. (2010). A probabilistic appearance representation and its application to surprise detection in cognitive robots, *IEEE Transactions on Autonomous Mental Development* 2(4): 267–281.

Meer, P. (1989). Stochastic image pyramids, *Computer Vision, Graphics, and Image Processing* 45(3): 269–294.

Montanvert, A., Meer, P. & Rosenfeld, A. (1991). Hierarchical image analysis using irregular tessellations, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13(4): 307–316.

Navalpakkam, V. & Itti, L. (2005). Modeling the influence of task on attention, *Vision Research* 45(2): 205–231.

Orabona, F., Metta, G. & Sandini, G. (2005). Object-based visual attention: a model for a behaving robot, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Patter Recognition (CVPR)*, p. 89.

Posner, M. I., Snyder, C. R. R. & Davidson, B. J. (1980). Attention and the detection of signals, *Journal of Experimental Psychology: General* 14(2): 160–174.

Rao, R. P. N. & Ballard, D. H. (1995a). An active vision architecture based on iconic representations, *Artificial Intelligence* 78: 461–505.

Rao, R. P. N. & Ballard, D. H. (1995b). Object indexing using an iconic sparse distributed memory, *Proc. the 5th Intl. Conf. Computer Vision*, pp. 24–31.

Rao, R. P. N., Zelinsky, G. J., Hayhoe, M. M. & Ballard, D. H. (2002). Eye movements in iconic visual search, *Vision Research* 42: 1447–1463.

Scholl, B. J. (2001). Objects and attention: the state of the art, *Cognition* 80(1-2): 1–46.

Sharon, E., Brandt, A. & Basri, R. (2000). Fast multiscale image segmentation, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 70–77.

Sharon, E., Galun, M., Sharon, D., Basri, R. & Brandt, A. (2006). Hierarchy and adaptivity in segmenting visual scenes, *Nature* 442: 810–813.

Shi, J. & Malik, J. (2000). Normalized cuts and image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(8): 888–905.

Specht, D. F. (1990). Probabilistic neural networks, *Neural Networks* 3(1): 109–118.

Sun, Y. (2008). A computer vision model for visual-object-based attention and eye movements, *Computer Vision and Image Understanding* 112(2): 126–142.

Sun, Y. & Fisher, R. (2003). Object-based visual attention for computer vision, *Artificial Intelligence* 146(1): 77–123.

Tipper, S. P., Howard, L. A. & Houghton, G. (1998). Action-based mechanisms of attention, *Philosophical Transactions: Biological Sciences* 353(1373): 1385–1393.

Treisman, A. M. & Gelade, G. (1980). A feature integration theory of attention, *Cognition Psychology* 12(1–2): 507–545.

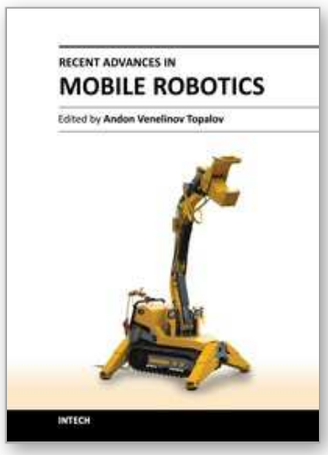
Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y., Davis, N. & Nuflo, F. (1995). Modelling visual attention via selective tuning, *Artificial Intelligence* 78: 282–299.

Walther, D., Rutishauser, U., Koch, C. & Perona, P. (2004). On the usefulness of attention for object recognition, *Workshop on Attention and Performance in Computational Vision at ECCV*, pp. 96–103.

Weng, J., McClelland, J., Pentland, A., Sporns, O., Stockman, I. & Thelen, E. (2001). Autonomous mental development by robots and animals, *Science* 291(5504): 599–600.

Wolfe, J. M. (1994). Guided search 2.0: A revised model of visual search, *Psychonomic Bulletin and Review* 1(2): 202–238.

Yu, Y., Mann, G. K. I. & Gosine, R. G. (2010). An object-based visual attention model for robotic applications, *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics* 40(5): 1398–1412.



Recent Advances in Mobile Robotics

Edited by Dr. Andon Topalov

ISBN 978-953-307-909-7
Hard cover, 452 pages
Publisher InTech
Published online 14, December, 2011
Published in print edition December, 2011

Mobile robots are the focus of a great deal of current research in robotics. Mobile robotics is a young, multidisciplinary field involving knowledge from many areas, including electrical, electronic and mechanical engineering, computer, cognitive and social sciences. Being engaged in the design of automated systems, it lies at the intersection of artificial intelligence, computational vision, and robotics. Thanks to the numerous researchers sharing their goals, visions and results within the community, mobile robotics is becoming a very rich and stimulating area. The book Recent Advances in Mobile Robotics addresses the topic by integrating contributions from many researchers around the globe. It emphasizes the computational methods of programming mobile robots, rather than the methods of constructing the hardware. Its content reflects different complementary aspects of theory and practice, which have recently taken place. We believe that it will serve as a valuable handbook to those who work in research and development of mobile robots.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen