Machine Vision and Applications
DOI 10.1007/s00138-009-0192-0

ORIGINAL PAPER

Cost-sensitive learning of top-down modulation for attentional control

Ali Borji · Majid N. Ahmadabadi · Babak N. Araabi

Received: 17 April 2008 / Revised: 27 December 2008 / Accepted: 20 January 2009

Abstract A biologically-inspired model of visual attention known as basic saliency model is biased for object detection. It is possible to make this model faster by inhibiting computation of features or scales, which are less important for detection of an object. To this end, we revise this model by implementing a new scale-wise surround inhibition. Each feature channel and scale is associated with a weight and a processing cost. Then a global optimization algorithm is used to find a weight vector with maximum detection rate and minimum processing cost. This allows achieving maximum object detection rate for real time tasks when maximum processing time is limited. A heuristic is also proposed for learning top-down spatial attention control to further limit the saliency computation. Comparing over five objects, our approach has 85.4 and 92.2% average detection rates with and without cost, respectively, which are above 80% of the basic saliency model. Our approach has 33.3 average processing cost compared with 52 processing cost of the basic model. We achieved lower average hit numbers compared with NVT but slightly higher than VOCUS attentional systems.

A. Borji (✉) · M. N. Ahmadabadi · B. N. Araabi
School of Cognitive Sciences,
Institute for Research in Fundamental Sciences,
Niavaran, P.O. Box 19395-5746, Tehran, Iran
e-mail: borji@ipm.ir

M. N. Ahmadabadi
e-mail: mnili@ut.ac.ir

B. N. Araabi
e-mail: araabi@ut.ac.ir

M. N. Ahmadabadi · B. N. Araabi
Control and Intelligent Processing Center of Excellence,
School of Electrical and Computer Engineering,
University of Tehran, Tehran, Iran

Keywords Selective perception · Top-down attention · Bottom-up attention · Basic saliency model · Object detection

1 Introduction

Both machine vision and biological vision systems are faced with the problem of processing enormous amount of visual information they receive at any given time. Attentional selection provides an efficient solution to this problem by proposing a small set of scene regions worthy further analysis to higher-level and cognitive processes like scene interpretation, object recognition, decision making, etc.

From a large body of literature in neurosciences and psychology, it is now known that attention is the process of selecting and gating visual information based on the saliency in the image itself (bottom-up) and on the prior knowledge about the scene (top-down). While bottom-up visual attention is solely determined by the basic and low-level physical characteristics of a scene-like luminance contrast, color contrast, orientation and motion—top-down attention on the other hand is influenced by the task demands, emotions, expectations, etc. Bottom-up component of the visual attention is mainly examined by the early visual areas of the brain like LGN, V1 and V2 [1,2]. Top-down attentional signals are largely derived from a network of parietal and frontal areas like frontal eye field (FEF), supplementary eye field (SEF) and lateral parietal cortex [3]. In daily life, these two mechanisms interact together to direct our attentional behaviors. Rather than acting in spatial domain [4,5], visual attention could also be directed to particular features [6] and objects [7].

As in biology, solutions in machine vision and robotics are limited in terms of processing huge amount of visual

Published online: 27 February 2009

sensory information, which is very time consuming. That is mainly because of the serial processing mechanisms used in the design of such creatures. This limitation necessitates engineering attentional control mechanisms in the brain of agents especially when they are supposed to act in real-world situations, which means to guarantee a limited response time. Some of the applications of attention in computer vision and robotics are object recognition, image compression, image matching, image segmentation, object tracking, active vision and human-robot interaction in addition to robot navigation and localization.

So far most experimental studies have addressed understanding bottom-up mechanisms of visual attention. That is probably because these mechanisms are mainly objective. On the other hand, top-down mechanisms show subject-to-subject variability, which makes them difficult to tackle [8]. As a result, computational studies have been concentrated more on modeling bottom-up mechanisms due to lack of abstract knowledge on top-down component of visual attention. In our work, we propose an approach for learning top-down attentional modulations without explicit mathematical formulation of impact of an object or its surrounding in saliency.

From a behavioral perspective, some of the human behaviors are learned by exposing them to a set of offline data. For example for learning to drive, a person familiarizes himself with traffic signs and their associated meanings and then uses this knowledge when doing real-time driving. Merging these two learning phases- offline learning of rules and signs and learning to apply them interactively and online—seems to be the best approach for doing this complicated task. According to this logic, our attention system could be considered as a basic component of a larger system, which provides top-down signals when requested by a higher cognitive layer.

We consider attention control as an optimization problem in which an artificial agent has to devote its limited processing power to the most relevant items or dimensions of the sensory space. The basic idea is that an agent is looking for an object of its interest in the scene and therefore has to bias its bottom-up attention system in order to detect that object *efficiently* and *fast*. We follow a data-driven approach for learning top-down attention, where top-down attentional signals are learned from a set of training images containing an object in clutter. This knowledge can later be used when an agent is doing a task online and needs to attend to different objects in different situations. The result of optimization will be ignoring some features (sensors) in addition to finding their relative importance in construction of a final saliency map.

Our approach is an extension of the basic saliency model [9], which is based on the idea of saliency map, an explicit two-dimensional topographical map that encodes stimulus conspicuity or saliency at every scene location. Saliency model is purely data-driven and simply selects some spatial locations without using any feedback mechanism or top-down gains. In particular, we introduce two contributions. First, the basic saliency model is revised in a way which allows selection and weighting of different scales. Then an evolutionary algorithm is used for biasing this model toward specific objects in cluttered scenes while considering the costs of operations of the saliency model. This allows an agent to use its certain limited processing resources efficiently in terms of achieving maximum object detection rate. In our second contribution, we propose a method to reduce saliency computation and therefore faster object search in natural scenes. It is based on the observation that, in some tasks objects appear in specific spatial locations with respect to the observer. For example in driving, traffic signs mostly appear in the right visual field of the driver. In addition to these, performances of our approach, basic saliency model as well as a general benchmarking approach for object detection are compared over disrupted images with several types of noises.

The rest of this paper is organized as follows. In Sect. 2, related works for learning top-down attention control and traffic sign detection are reviewed. Basics of our method for learning top-down feature-based and spatial attention control are presented in Sect. 3. Experiments and results are shown in Sect. 4. Section 5 discusses the results and finally, Sect. 6 summarizes and concludes the paper.

2 Related works

A broad range of modeling works are reported in the literature of visual attention, however little research is reported on learning and development of attention control. In this section we review some studies which are directly related to ours, mainly those focused on learning aspects of visual attention control. In order to situate our work among previous works on traffic sign detection, some successful studies from this domain are also reviewed.

2.1 Learning top-down attention control

Evidence toward the idea that attention could be learned by biological mechanisms is proposed in [10]. In a behavioral task, authors showed that a short-term memory system is responsible for rapid deployment of visual attention. They suggest that fast and transient component of visual attention is flexible and capable of learning simple relationships [11] and is dependent on the previous experiences of the subjects [12]. Human subjects were supposed to answer a question about a feature of a specific visual item in a synthetic search array. Subjects had lower reaction times when this feature remained the same throughout successive trials [13].

In [14], it is postulated that human eyes move and successively fixate at the most salient parts of an image during visual perception and recognition. These parts are then processed with the highest resolution. Simultaneously, the mechanism of visual attention chooses the next eye position using information extracted from the retinal periphery [15]. Eye movement behavior has been shown to be task-specific and context-dependent in humans [8]. In a behavioral task, human subjects were asked a question about a scene presented to them and their saccadic eye movements were recorded. Depending on the question, subjects had different eye movement patterns. For example when asking to judge about age of persons in a scene, eye movements were mainly focused on faces. This experiment points toward another evidence that top-down attention can be learned.

One of the outstanding works on visual attention, known as the basic saliency model, is proposed by Itti et al. [9] which is an extension and implementation of an earlier model of visual attention by Koch and Ullman [16]. Simplicity and little processing time are two main advantages of this model. It has been used to explain behavioral data on simple synthetic and static search arrays and also dynamic natural stimuli like movies and games [17]. To add top-down capabilities to this model, a task-driven model of visual attention is proposed in [18]. Given a task definition in the form of keywords, this model first determines and stores task-relevant entities in working memory using prior knowledge stored in a long-term memory. It then attempts to detect the most relevant entity by biasing its visual attention system with the entity's learned low-level features. It attends to the most salient location in the scene and attempts to recognize the attended object through hierarchical matching against object representations stored in the long-term memory. It updates its working memory with the task relevance of the recognized entity and updates a topographic task relevance map with the location and relevance of that entity. In this study, we aim to build our top-down attention system upon the basic saliency model and bias it for detection of different objects in natural scenes.

In [19], Frintrop et al. have introduced a new computational attention system known as VOCUS for efficient and robust detection of regions of interest in images. In their approach, the selection of an image region is determined by two kinds of influences: bottom-up and top-down cues. Bottom-up cues are determined by local contrasts and by uniqueness of a feature. Top-down cues depend on the features of a pre-specified target. Bottom-up and top-down saliency maps are then linearly added and weighted to form a final saliency map. They have shown that VOCUS is robust and applicable to real-world scenes.

Basic saliency model concentrates on computing bottom-up attention. Recently, Navalpakkam and Itti [20] have introduced a newer version of the basic model by adding top-down capabilities to it. The idea is to learn feature values of a target object from a training image in which the target is indicated by a binary mask. By considering the target region as well as a region in its close surrounding, their system learns feature values from different feature maps. During object detection, this feature vector is used to bias the feature maps by multiplying each map with the corresponding weight. Thereby, exciting and inhibiting as well as bottom-up and top-down cues are mixed and directly fused into the resulting saliency map.

In contrast to the above mentioned two techniques [19,20], instead of only finding the appropriate weights, we also incorporate processing costs of the feature channels to force an optimization process to choose the feature vectors with high detection rate and low cost. Basic saliency model does not allow scale selection because center-surround inhibition in this model is implemented by subtraction of scales from each other. We revise the basic model by implementing surround inhibition in each scale independent of others. This not only allows scale weighing but also allows selection among scales for saliency detection. Inhibition of those scales, which are not important in detection of an object results in faster object detection and less computation. Furthermore associating costs to operations of the basic saliency model allows optimal use of computational resources of the agent. For example when an agent has a certain amount of computation power, it must efficiently select/weight features and scales of the saliency model in order to achieve the maximum detection rate. Our approach in this paper provides such a capability.

2.2 Traffic sign detection and recognition

Traffic sign recognition (TSR) can be considered as part of the bigger problem of autonomous driving. An autonomous driver system relies on vision-based recognition of surrounding area in order to make driving system function as the feedback provider for control of steering wheel, accelerator, brake, etc. Besides the application of TSR in autonomous vehicles, it can also serve as an assistant driver to notify the driver about approaching a traffic sign or his risky behavior (like driving above speed limit). Driving is the best example of a complex task, which involves many cognitive behaviors and most importantly attention. Due to this, we consider visual attention as a part of a TSR system for fast detection of traffic signs.

Many researchers have developed various techniques for automatic road traffic sign recognition. Regarding the detection problem, different approaches have been proposed. A few approaches rely solely on gray-scale data. Gavrila [21] employs a template-based approach in combination with a distance transform. Barnes and Zelinsky [22] utilize a measure of *radial symmetry* and apply it as a pre-segmentation within their framework. Since radial symmetry corresponds to a simplified (i.e., fast) circular Hough transform, it is particularly applicable for detecting possible occurrences of circular signs. The majority of recently published sign detection approaches make use of color information [23–25]. They share a common two-step strategy. First, a pre-segmentation is employed by a thresholding operation on a color representation. Some authors perform this directly in RGB space, others apply its linear or nonlinear transformations. Subsequently, a final detection decision is obtained from shape based features like corners or edge features, applied only to the pre-segmented regions. The most discriminating colors for traffic signs include red, orange, yellow, green, blue, violet, brown and achromatic [25]. A joint treatment of color and shape has been proposed by Fang et al. [26]. The authors compute a feature map of the entire image frame, based on color and gradient information, while incorporating a geometry model of signs. Detection of traffic signs in only a single image has three problems [27]: (1) information about positions and size of traffic signs, which is useful to reduce the computation time, is not available (2) it is difficult to detect a traffic sign correctly when temporary occlusion occurs and (3) correctness of the detection is hard to verify. In order to handle these issues tracking techniques have been developed for traffic sign detection. In [27], image sequences are utilized for recognition of traffic signs.

Discriminating shape and colors of traffic signs make them easily recognizable by humans. Same factors bring the idea of applying the basic saliency model for detection of traffic signs, which we follow in this work. Not only we analyze the capability of the saliency model for detecting traffic signs, but also we evaluate its efficiency for natural object detection.

The main reason why we propose the basic saliency model of visual attention as a subsystem of a typical TSR system for detection of traffic signs is because of its fast computation time. This does not necessarily mean that it could outperform above mentioned methods for detection of traffic signs. There are two reasons for this. First, basic saliency model is based on abstract biological findings from human vision and is designed in a way to be fast like human visual attention. It is a general purpose system for detection of salient image regions. Second, basic saliency model in its current form does not consider structure and shape of objects. This weakness makes it inferior (bus faster) to single purpose solutions for detection of specific objects like faces or traffic signs. Adding learning capabilities to the basic saliency model for enabling it to also consider shape of different objects demands more research and goes beyond the scope of this paper.

3 Learning top-down attentional modulations

In this section we present our goal-directed visual attention system. Top-down influences play an important role in human visual attention rather than bottom-up cues. Top-down sources of visual attention are mainly determined by the experiences, motivations, emotions, expectations and goals. For example a hungry person is focused on foods.

Saliency model without biasing selects a spatial location in a feed-forward manner. Saliencies in feature dimensions—intensity, orientation and color—are computed independently and are fused into a single final saliency map. The top-down extension in our model includes a search phase to learn feature weights to determine which features to enhance and which ones to inhibit. The weighted features contribute to modulate the bottom-up saliency map highlighting regions with target-relevant features. Instead of only finding the appropriate weights, we also incorporate processing costs of the feature channels to force the optimization process to choose the feature vectors with high detection rate and low cost. When an agent has a certain amount of processing resources it could bypass computation of irrelevant feature channels or scales of the basic saliency model. This is actually a constrained optimization problem, which we solve by global optimization techniques.

3.1 Revised saliency-based model of visual attention

To furnish the basic saliency model for our purpose, i.e biasing it for object detection, a Gaussian pyramid [28] with six scales ($s_0, \ldots, s_5$) is built in each feature channel of the image by successively filtering the image with a Gaussian low-pass filter (GLF) that is then subsampled, i.e s_{i+1} has half the width and height of s_i. Three features are considered; intensity, orientation and color. A surround inhibition operation is then applied to all scales of each pyramid to enhance those parts of the image, which are different from their surroundings. To implement surround inhibition, basic saliency model [9] subtracts coarser scales from finer ones. Since we would like to select scales rather than weighting them, we dissociate scales and then apply the surround inhibition over each scale separately. This is not possible in the basic model because scales are dependent to each other and hence cannot be inhibited or selected. Surround modulated images in each scale of a pyramid are upsampled to a predetermined scale (here the largest scale, s_0) and are then added. To accomplish surround inhibition (SI), we designed a nonlinear filter, which compares the similarity of each pixel with the average of its surrounding window and then inhibits the center pixel as:

$$\begin{aligned}&Im' = \mathrm{SI}(Im),\\&Im'(x,y) = \max(0, Im(x,y) - \mathrm{mean}(\mathrm{surround}(Im(x,y)))),\\&\qquad \forall x, y \in Im \qquad (1)\end{aligned}$$

where surround is a spatial $n \times n$ mask around pixel *Im(x, y)*. $Im'(x, y)$ is the new value of the pixel. Figure 1 demonstrates

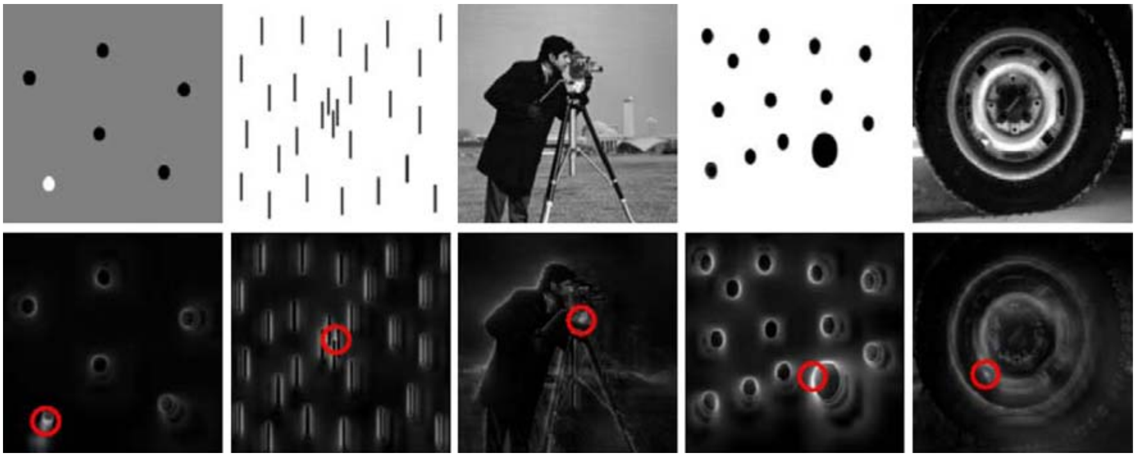

Fig. 1 Surround inhibition operation using (1). *Top row* shows five test images and *bottom row* shows the detected salient regions with surround window sizes (n) from *left* to *right* as 7, 7, 3, 7 and 5. *Red circles* illustrate the most salient locations

application of the surround inhibition operation on some test images. Here, all scales of the intensity pyramid are surround inhibited and then added to form the saliency map at the intensity channel. As Fig. 1 shows, this operation has resulted in marking the salient areas with respect to the rest of each image.

We used a MATLAB® implementation of the basic saliency model and revised it for our purpose [30]. The whole revised saliency model is shown in Fig. 2.

The input image to the system is decomposed into three-feature channels: Intensity (I), Color (C) and Orientation (O). Color channels are calculated as follows. If r, g and b are the red, green and blue dimensions in RGB color space, then $I = (r + g + b)/3$, $R = r - (g + b)/2$, $G = g - (r + b)/2$, $B = b - (r + g)/2$, and $Y = r + g - 2(|r - g| + b)$ (negative values are set to zero). Local orientations (O_θ) are obtained by applying Gabor filters to the images in the intensity pyramid I. These operations are shown in (2). P_s is the feature at scale s. P could be intensity (I), Red (R), Green (G), Blue (B), Yellow (Y) or orientation (O). $O_{\theta,s}$ is the orientation channel at orientation θ and scale s.

$$\begin{aligned} F_{I,s} &= SI(I_s) \\ F_{RG,s} &= SI(R_s - G_s) \\ F_{BY,s} &= SI(B_s - Y_s) \\ F_{\theta,s} &= SI(O_{\theta,s}) \end{aligned} \tag{2}$$

In (2), $F_{I,s}$, $F_{RG,s}$, $F_{BY,s}$ and $F_{\theta,s}$ are the intensity, red/green, yellow/blue and orientation channels in scale s, respectively. SI is the surround inhabitation operation in (1). These feature maps are summed over and sums are normalized again:

$$F_l = N\left(\sum_s (s\omega)_s . F_{l,s}\right) \quad \text{with} \quad l \in L_I \cup L_C \cup L_O$$
$$\text{and } L_I = \{I\},\ L_C = \{RG, BY\},\ L_O = \{0°, 45°, 90°, 135°\} \tag{3}$$

where $(s\omega)_s$ is the weight of scale s. $N(.)$ is an iterative, nonlinear normalization operator, simulating local competitions between neighboring salient locations [29]. In each feature channel, feature dimensions contribute to the conspicuity maps by weighting and normalizing once again (4).

$$C_p = N\left(\sum_{l \in L_p} (d\omega)_p \cdot F_l\right), \quad p \in \{I, C, O\} \tag{4}$$

Variable $(d\omega)_p$ in (4) determines weight of a dimension within feature channel p. All conspicuity maps are weighted and combined at this stage into a final saliency map (5).

$$SM = \sum_k (c\omega)_k \cdot C_k, \quad k \in \{I, C, O\} \tag{5}$$

$(c\omega)_k$ in (5) weights the influences of feature channels in the final saliency map. The locations in the saliency map compete for the highest saliency value by means of a *Winner-Take-All* (WTA) network of integrate and fire neurons [16]. The winning location of this process is attended to and the saliency map is inhibited at this location. Continuing WTA competition, next most salient location is attended to and so on to form a scanpath of successive overt attentions.

3.2 Offline learning of top-down attentional modulations

Real-world systems have limited amount of computational resources due to serial processing mechanisms used in their designs. Therefore it is necessary to optimize use of these

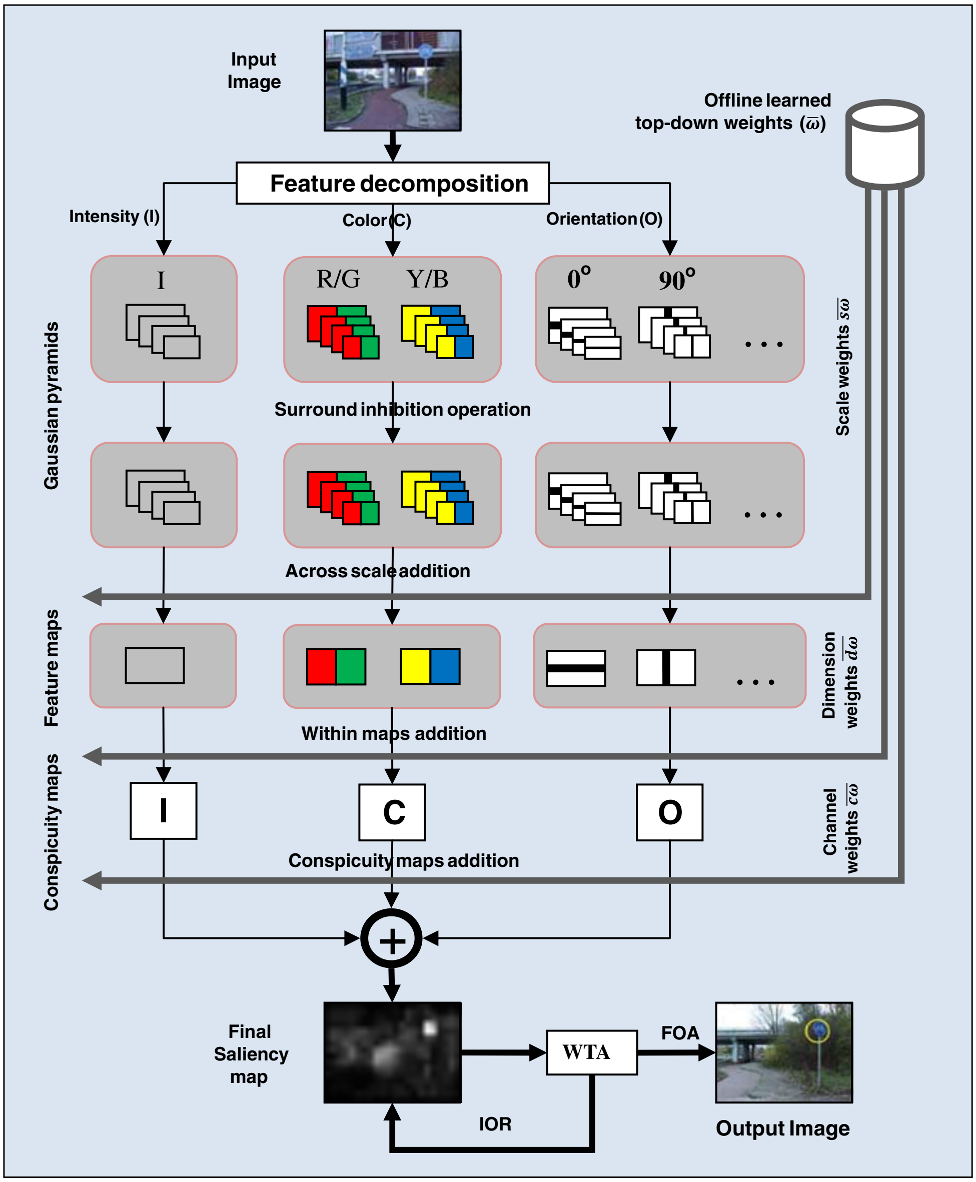

Fig. 2 Proposed top-down saliency model. Input image is decomposed to several feature channels and dimensions within each feature channel. Here, one dimension for intensity channel, two red/green and yellow/blue dimensions for color channel and four dimensions for orientation channel are used. Image pyramids are built in each feature dimension by successively low-pass filtering and subsampling the image. Surround inhibition operation (1) is applied to all scales of a channel. In the next step, maps across scales are weighted and summed to form feature maps (2, 3). Then, feature maps are normalized, weighted and summed again to form conspicuity maps (4). Final saliency map is derived by normalizing, weighting and summing these conspicuity maps (5). Maximums at the final map determine the focus of attention (FOA). After a time constant, the most salient location is inhibited to allow other locations grab attention (IOR). Contributions of scales in dimensions, dimensions in feature channels and feature channels in the final saliency map are determined by a learned weight vector ($\bar{\omega}$)

resources. For the agent to make optimum use of its processing resources, in addition to a weight, a cost is also associated to each feature channel and image resolution. Weight and cost vectors are represented by $\bar{\omega}$ and $\bar{C}$, respectively in (6). Weight vector determines weights of feature channels, dimensions within channels and also scales in image pyramids respectively. Elements in the cost vector correspond to the associated feature or scale determined by the weight vector.

$$\bar{\omega} = \left(\overline{c\omega}, \overline{d\omega}, \overline{s\omega}\right), |\bar{\omega}| = 16, |\overline{c\omega}|=3, |\overline{d\omega}|=7, |\overline{s\omega}|=6$$
$$\overline{d\omega}=(\overline{d\omega I}, \overline{d\omega C}, \overline{d\omega O}), |\overline{d\omega I}|=1, |\overline{d\omega C}|=2, |\overline{d\omega O}|=4$$
$$\bar{C} = (c_1, c_2, \ldots, c_{16}) \tag{6}$$

In (6) $\overline{c\omega}$, $\overline{d\omega}$, $\overline{s\omega}$, are weight vectors for feature channels, dimensions within channels and scales, respectively. $\overline{d\omega I}$, $\overline{d\omega C}$, $\overline{d\omega O}$ are weight vectors for intensity, color (red/green and yellow/blue) and orientation (0°, 45°, 90° and 135°) dimensions. Our aim is to find a vector of weights ($\bar{\omega}$) determining the contributions of features and image resolutions to detect a specific object in a set of images while taking into account their costs ($\bar{C}$). Thus, the optimum weight vector must satisfy these two objectives (1) It must enable the saliency model to detect an object of interest correctly (maximum detection rate) and (2) Its associated feature vector, must have the least computation (minimum processing cost). For this purpose, we follow a data-driven approach. First, optimal weight vector is sought to satisfy the above two objectives over a training image dataset and is then evaluated over a separate test set of images. Assume that training set contains M images with an object of interest tagged in them as:

$$T = \{Im_1, Im_2, \ldots Im_M\} \tag{7}$$

An example fitness function, which satisfies the above mentioned two objectives is shown in (8).

$$f(\bar{\omega}) = \frac{1}{M}\left(\sum_{i=1}^{M} \text{norm}(\text{Saliency}(Im_i, r(\bar{\omega})) - t_i)\right)(u(\bar{\omega}) \cdot \bar{C})$$
$$r(\omega_i) = \begin{cases} \omega_i, & \omega_i > \alpha \\ 0, & \text{otherwise} \end{cases}, \quad u(\omega_i) = \begin{cases} 1, & \omega_i > \alpha \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

In (8), M is the size of the train set, norm(.) is the Euclidean distance between two points in an image. Saliency is a function, which takes an image and a vector of weights as input and determines the most salient location in the image. t_i is the location of the target object in the ith image. $u(.)$ is the step function and is 1 when a feature channel or resolution is used. $r(.)$ is the ramp function and zeros its input when it is smaller than a threshold. Operator $\cdot$ is the dot product. Since we are going to compare the saliency detection with and without costs, results are also reported using the fitness function of (9) when costs are ignored and the single objective is to maximize the detection rate.

$$f(\bar{\omega}) = \frac{1}{M}\left(\sum_{i=1}^{M} \text{norm}(\text{Saliency}(Im_i, r(\bar{\omega})) - t_i)\right) \tag{9}$$

We also consider a case when the agent has a limited computational power and has to select those feature channels or scales of the basic saliency model when the accumulated cost does not exceed a cost limit (Q). This is a constrained optimization problem in which goal is to maximize detection rate but with a constraint which is the cost limit:

$$f(\bar{\omega}) = \frac{1}{M}\left(\sum_{i=1}^{M} \text{norm}(\text{Saliency}(Im_i, r(\bar{\omega})) - t_i)\right)$$
$$u(\bar{\omega}) \cdot \bar{C} \le Q \tag{10}$$

Note that, a lower fitness value for the above functions means that it has better performance. For minimizing fitness functions, an algorithm known as comprehensive learning particle swarm optimization (CLPSO [31]) is used. CLPSO is simple, easy to implement and has been applied to a wide range of optimization problems. Its fast convergence and better performance over multi modal functions are the reasons why we chose this optimization algorithm. First, particles of CLPSO (each particle is a 16D weight vector) are randomly initialized in the range [0 6]. Salient locations are detected by applying each particle to the saliency model to calculate its fitness. Then, through CLPSO operations, particles are evolved in order to minimize the fitness values. In cost-limited case (10) when an individual violated the constraint condition, its fitness was assigned a very large positive number. Table 1 shows parameters of CLPSO used in the experiments.

Table 1 Parameters of CLPSO used in experiments (refer to [31] to learn about parameters)

Parameter	Value
Particles	300
Max iterations	40
Dimensions	16
$[\min(x_i)\max(x_i)]$	[0 6]
Refreshing gap	5
$[\omega\, c\, p_c\, v_{\max}]$	[0.9 1.5 0.3 5]

4 Experiments and results

Humans have the capability to attend to features or dimensions of perceptual space or could select the spatial locations, which usually contain objects [7]. The same capability is highly desired for artificial systems. In this section by means of four experiments we show how our algorithm can be used for both feature-based and spatial attention control.

4.1 Learning top-down feature based attention control

Three experiments demonstrate the strength of our algorithm for feature-based attention control. First two experiments are pop-out and conjunction search tasks [32] and object detection in natural scenes. Search performance (percentage of detection rate), average hit number and cost of our method are compared with the basic saliency model in both without-cost (9) and with-cost (8) cases. In the third experiment, performance of our method is compared with the basic saliency model as well as template matching (TM) approach over noisy images.

In with-cost case, feature costs were defined based on relative computational costs of feature channels and image resolutions as $\bar{C} = [3, 1, 4, 3, 3, 1, 4, 4, 4, 4, 6, 5, 4, 3, 2, 1]$. For example, a color channel needs more computation than an intensity channel but less compared to an orientation channel. Or computation of surround inhibition in scale s_0 is more expensive than those of other scales in the Gaussian pyramid.

Experiment I. Pop-out and conjunction search tasks In a pop-out or conjunction search task, a search array is presented to the system. It has to determine, which item in the search array is different from other items. Then its reaction time is measured. Items in the search array differ in one dimension only in a pop-out task, while in a conjunction search task, items differ in more than one dimension, which makes the task harder. Psychological data have shown that reaction times (RT's) of human subjects remain constant with increasing the number of items in the search array. In contrast, in a conjunction search task, reaction times increase linearly with size of the search array.

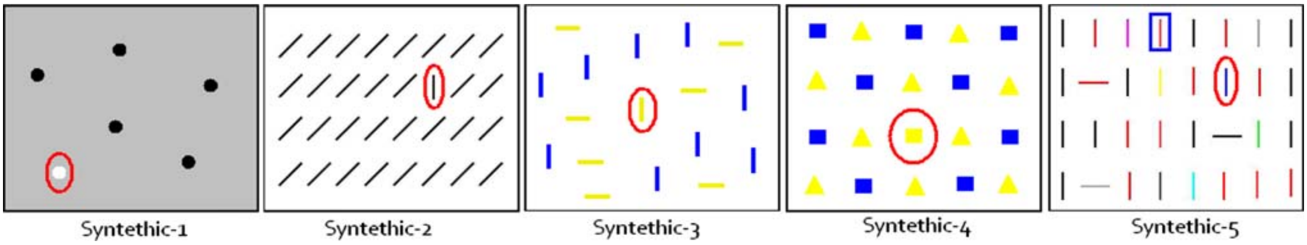

Fig. 3 Synthetic search arrays used in Experiment I. Two left arrays are pop-out and three right ones are conjunction search tasks. Target item is shown with *red ellipse*. *Blue rectangle* in the fifth array is the first saccade of the basic saliency model

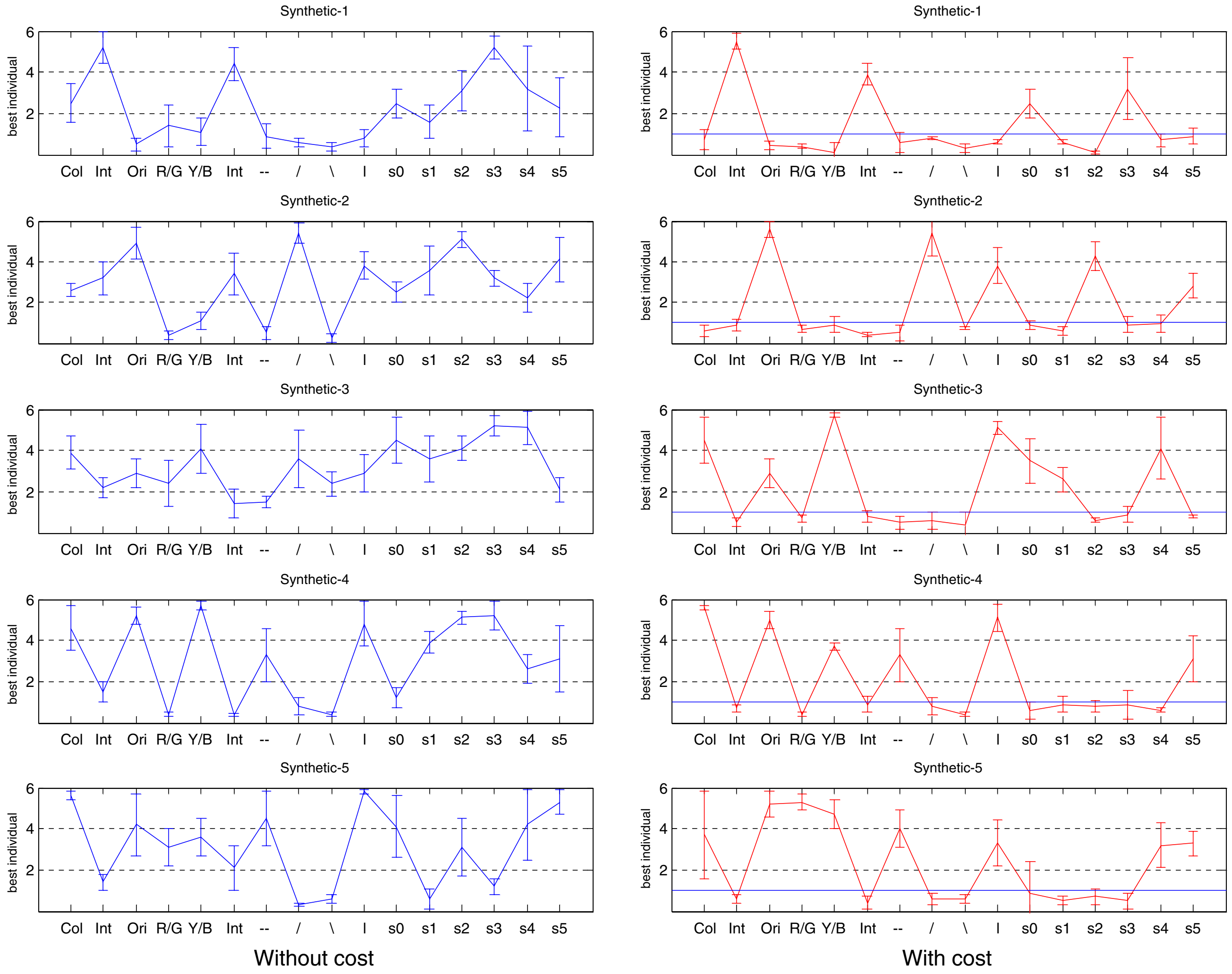

Fig. 4 Learned weights after CLPSO convergence over synthetic search arrays. *Left column* shows the weights learned in without-cost case and right column shows with-cost case. *Blue horizontal lines* in the *right column* show lines $\alpha = 1$ (threshold parameter in (8)). Employed features are: Color, Intensity, Orientation, Red/Green and Yellow/Blue dimensions within color channel, a single intensity dimension, four orientation dimensions within orientation channel and six scales

Revised saliency model was trained using CLPSO for known saliency detection over two pop-out and three conjunction search arrays shown in Fig. 3. Figure 4 shows biasing weights over the synthetic search arrays in both without-cost and with-cost cases after CLPSO convergence. CLPSO is trained five times and averages are shown. In this experiment, size of the surround inhibition window was 5. In all search arrays target item was successfully detected in the first fixation.

As Fig. 4 shows, in the first search array with white dot among black dots, intensity channel has the highest weight in both cases. In with-cost case of this array, only intensity

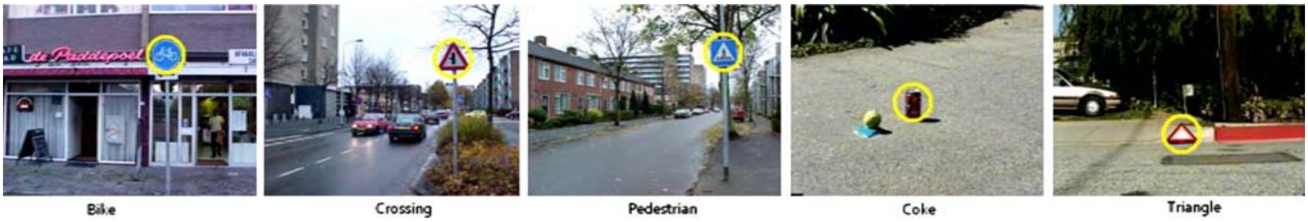

Fig. 5 Sample objects used in Experiment II. From *left* to *right* bike, crossing, pedestrian, coke and triangle. Target is shown by the *yellow circle*

channel and scales s_0 and s_3 are selected resulting in mean total cost of 11. Since orientation is not important in saliency of the target it is ignored by CLPSO. In the second array with vertical bar among oblique bars, in without-cost case, orientation channel and two dimensions (45° and 90°) have the highest weights. In with-cost case other channels except orientation are suppressed to reduce cost from 41 to 17. Intensity channel is ignored because intensities of target and distracters are the same. Since the target in the third array is the yellow bar, yellow/blue color dimension have the highest weight. For this array scales s_0–s_4 are more important. In the fourth array, color channel and yellow/blue dimension have the highest weights. In with-cost of this array only yellow/blue, 0° and 90° dimensions are selected. While in the first two pop-out search arrays, the basic saliency model can also detect the target, in the conjunction search arrays, CLPSO was trained to selectively attend to a target, which is not necessarily the attended item by the basic saliency model. For example, while in the last synthetic image, the basic saliency model selects the bar in the rectangle, we purposefully chose the blue bar to become salient. Since there are several colors available in this array, both dimensions within the color channel have got high weights. Weights for intensity channel and orientations (45° and 135°) are very low. In with-cost case of this array, intensity, orientation and scales (s_0 to s_3), which are not discriminating the target object are not selected by the evolutionary process.

Experiment II. Natural object and traffic sign detection Proposed method was also evaluated for natural object and traffic sign detection on cluttered scenes.[1] We used three traffic signs (bike, crossing and pedestrian) and two objects (coke and triangle). Number of images for bike, crossing, pedestrian, triangle and coke were 70, 45, 55, 69 and 42, respectively. Sizes of images were 360 × 270 pixels. Figure 5 illustrates sample signs and objects in natural scenes.

CLPSO was trained over ten random images for each object and then the best weight vector was tested over the remaining images of that object. Results are reported over five runs with random train images. Figure 6 shows fitness values for both without-cost and with-cost cases during CLPSO convergence over training sets. Derived weight vectors in both cases are shown in Fig. 7. Window size of surround inhibition was seven.

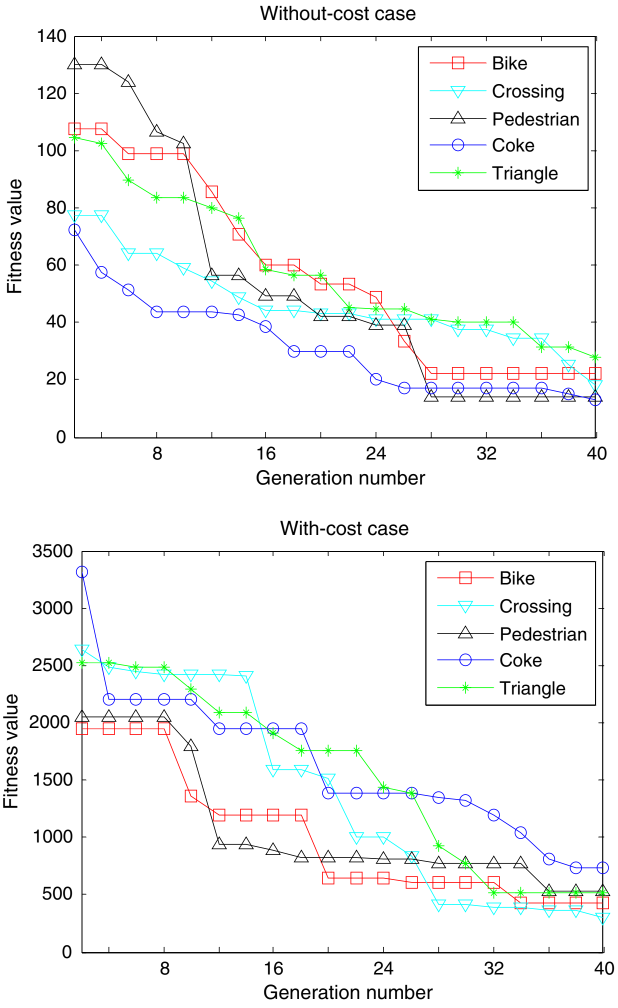

Fig. 6 CLPSO convergence for traffic signs and natural objects in Experiment II in both without-cost (*top*) and with-cost (*bottom*) cases

For bike, in without-cost case, color (yellow/blue) and orientation (45° and 135°) channels have the highest weights. Middle scales (s_1–s_3) are more informative for detection of this object. In with-cost case, the optimization process has

[1] Images for this experiment were selected from databases available at http://ilab.usc.edu/imgdbs and http://www.cs.rug.nl/~imaging/databases/traffic_sign_database/traffic_sign_database.html.

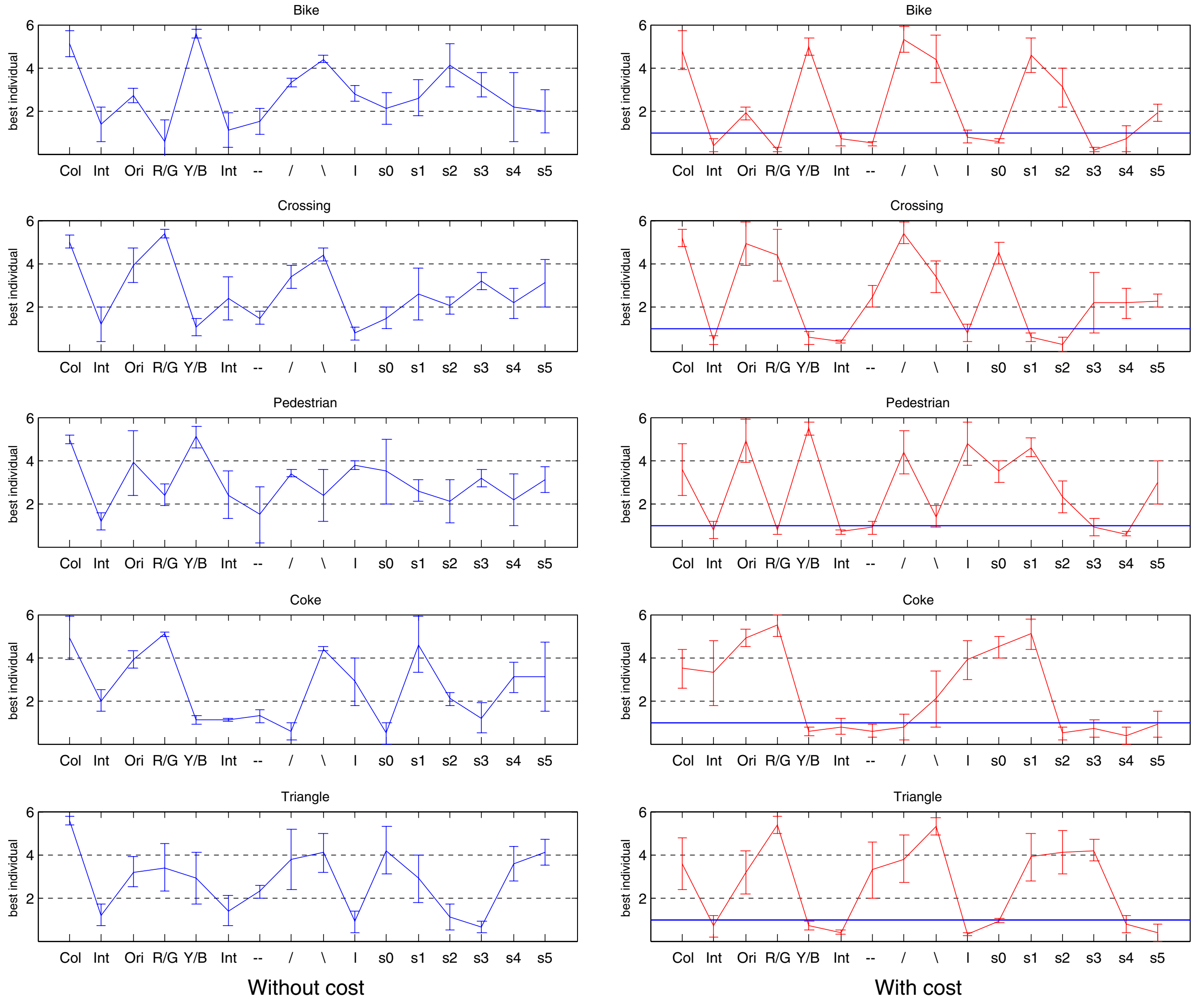

Fig. 7 Learned weights after CLPSO convergence over traffic signs and objects. *Left column* shows the weights learned in without-cost case and *right column* in with-cost case. *Blue horizontal line* in *right column* shows the line $\alpha = 1$ (threshold parameter in (8)). *Values* below this line are considered zero to evaluate the detection rates

decreased the contribution of intensity and horizontal orientation. For crossing, again color channel is selected in without-cost case. For this sign, red/green channel has higher weight. Orientation dimensions (0°, 45° and 135°), which appear in shape of the crossing sign have higher weights. Since almost in all images of this sign, triangle is toward up, these orientations are stable features and have got higher weights than other orientation dimensions. In with-cost case of this sign, color and orientation channels have survived in the evolutionary process. In this case, the largest scale s_0 has the highest weight. For pedestrian sign, color (yellow/blue) and orientation channels are selected in both cases. Important orientations for this sign are 45° and 90°.

For the coke object, color (red/green) channel, orientation 135° and scale s_1 have higher weights in without-cost case. In both cases orientations 90° and 135° are important for detection of this object. For triangle object as in crossing sign in both cases, 0°, 45° and 135° orientations have higher values again showing importance of these features in discrimination of triangular objects. Among color channels, red/green dimension has the highest weight.

Tables 2 and 3 show the average values of detection rates and average hit numbers using fitness functions in (8) and (9). An object was considered detected if a salient point was generated in a vicinity of 30 pixels around t_i. For detection of an object rather than the most salient point, 2 other locations generated by WTA were also considered. Hit number of an image Im_i for a known target t_i is the rank of the focus that hits the target in order of saliency. For example, if the second focus is on the target then its hit number is 2. Images

Table 2 Detection performances of biased saliency model in without-cost case and the basic saliency model over natural objects and traffic signs with max fixations equal to 3

Target	Biased saliency model						Basic saliency model	
	Train			Test			Detection rate %	Avg. hit number
	Detection rate %	Avg. hit number	Fitness	# Test images	Detection rate %	Avg. hit number		
Bike	92.3 (2.1)	1.4 (0.4)	22.3 (9.2)	60	90.2 (2)	1.6 (0.1)	81.8 (0.6)	2.2 (0.2)
Crossing	96.7 (1.2)	1.5 (0.7)	18.1 (5.4)	35	93.8 (0.9)	1.5 (0.2)	78.2 (1.4)	2.5 (0.3)
Pedestrian	98.2 (1)	1.2 (0.2)	14.4 (6.1)	45	94.2 (1.1)	1.3 (0.7)	83.3 (1)	1.7 (0.1)
Coke	95.2 (1.4)	1.3 (0.3)	13.2 (8)	59	92.2 (2)	1.5 (0.5)	80.9 (0.4)	1.9 (0.5)
Triangle	92.5 (2.3)	1.7 (0.9)	27.8 (11.4)	32	91 (1.6)	1.8 (0.4)	76.5 (0.8)	2.2 (0.2)

Results are averaged over 5 runs with random training and test sets. Performance of the basic saliency model is over the test sets. Numbers in parentheses are standard deviations

Table 3 Detection performance over natural objects and traffic signs in with-cost case with max fixations equal to 3

Target	Train			Test			
	Detection rate %	Avg. hit number	Fitness	# Test images	Detection rate %	Avg. hit number	Computation cost
Bike	87.8 (1.2)	1.7 (0.3)	421.9 (25.6)	60	85.8 (1.5)	1.8 (0.2)	25.5 (4.2)
Crossing	84.2 (1.9)	1.5 (0.7)	302.1 (47.1)	35	78.2 (2.1)	2 (0.3)	40.5 (5.8)
Pedestrian	91.6 (1)	1.5 (0.4)	531.9 (54.3)	45	90.6 (2)	1.7 (0.4)	35.6 (5.2)
Coke	87.3 (2.7)	1.8 (0.6)	730 (34.4)	59	87.1 (1.1)	1.9 (0.1)	32.2 (7.3)
Triangle	89.6 (2.1)	1.4 (0.2)	512.2 (27.7)	32	85.6 (1.2)	2.1 (0.2)	32.7 (6.1)

with targets not detected in max fixations are not included in the averages. The average hit number for an image set is the mean of hit numbers of all images. A search method with lower average hit number is favored.

The basic bottom-up saliency model has always total cost of 52, since it uses all the channels and resolutions. On the other hand, biased model has the average cost of 33.3 over signs and objects, which is less than the cost of the basic saliency model. Average detection rate with 3 hits of the basic saliency model is 80% while we achieved 92.2% without cost and 85.4% with cost. Results prove that biasing leads to higher recognition rate and lower cost than the basic saliency method. Average hit numbers of our approach are 1.54 (without-cost) and 1.9 (with-cost) while the basic saliency model has average hit number of 2.1. Performance of our model without biasing is nearly the same as the basic saliency model. Variance in the costs shows that cost is sensitive to train and test data.

Figure 8 shows detection rates of cost-limited case (10) for several values of cost. As the cost limit increases, detection rate also increases. In this case the best individual learned with a certain amount of cost is applied to the test set. This figure helps an agent to assign its processing power to reach the accuracy level it needs. For example in order to reach 80% for bike detection, the agent should at least have 20 computational resources or costs. Or with 40 processing cost the agent could not achieve more than 89% for detection of coke using the biased saliency model. For analysis of the data difference in fitness functions and standard deviations should be noted.

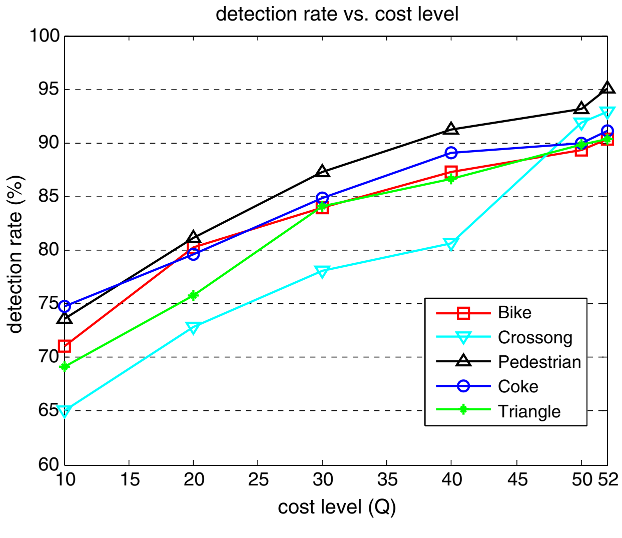

Fig. 8 Mean detection of cost-limited case for traffic signs and objects over test sets

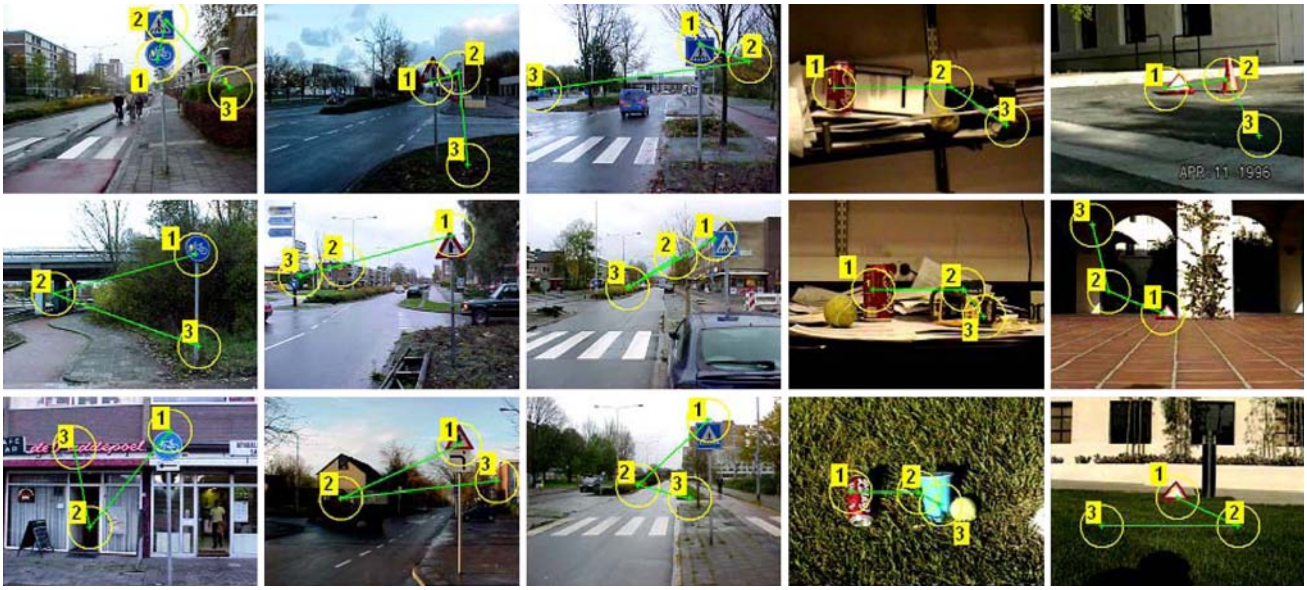

Fig. 9 Three first attended locations of the biased saliency model over traffic signs and the objects. *Numbers* indicate the order of attended locations. *Columns* from *left* to *right* are for bike, crossing, pedestrian, coke and triangle objects

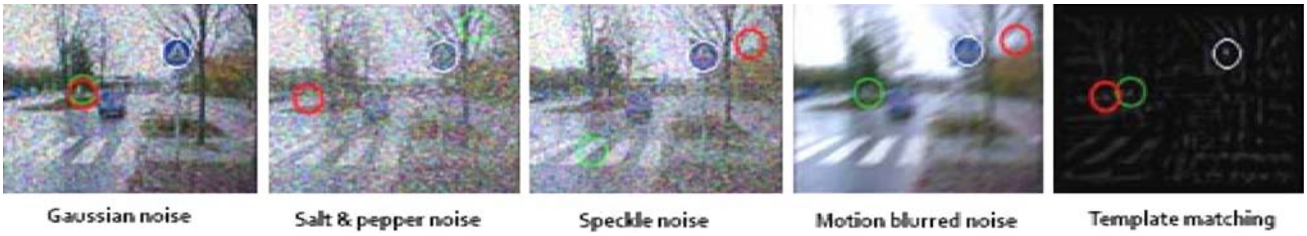

Fig. 10 Pedestrian sign detection over noisy images with biased saliency model. From left to right: Gaussian noise ($\mu = 0, \sigma = 0.5$), salt & pepper noise ($d = 0.6$), speckle noise with density $d = 0.5$, motion blurred with a 20×30 window and correlation map over original noiseless image in template matching. White circle illustrates the most salient location and other 2 circles show less similar ones

Three first salient points generated by the biased saliency model (using the best individual in the final population in without-cost case) over three sample images of each object are shown in Fig. 9.

Experiment III. Detection performance over disrupted images In this experiment, we compare the biased saliency model in without-cost case against basic saliency model and template matching, which is a basic benchmarking method for object detection over disrupted images with several types of noises. In template matching, a window containing an instance of an object is slided over the image and correlation (or convolution) of each image region with the object is calculated. Template matching is very sensitive to image distortions like in-plane rotation and also to template selection. Gaussian, salt and pepper, speckle and motion blurred noises were selected because they simulate rainy and snowy weathers and movements in driving situations.

Achieved weights after CLPSO training in Experiment II are used in this experiment over noisy test images. Three salient locations were proposed by each method as the most probable locations containing signs. Figure 10 illustrates a sample image under typical noises and 3 locations proposed by the biased saliency model. Table 4 compares detection rates of the biased saliency model with template matching and basic saliency model.

As results in Table 4 show biased saliency model performs better than the basic saliency model over all noises and have near the same computation time as it. While template matching has lower detection rates compared with biased saliency model, it is very sensitive to Gaussian noise and is more stable over other noises. However its computation time is much more than other two approaches. Template matching needs 3.5 times computation time more than the basic saliency model on average (3 times more than biased saliency model) on a computer with 1.4 GHz Intel cpu and 512 MB ram memory.

4.2 Learning top-down spatial attention control

In many tasks, humans are biased (from their past experiences) toward specific spatial locations of their environment.

Table 4 Mean detection rates and computation times of traffic signs over test images with 5 runs

Method	Without noise	Gaussian ($\mu = 0, \sigma = 0.5$)	Salt and pepper ($d = 0.6$)	Speckle ($d = 0.5$)	Motion blurring, window size (20× 30)	Computation time (ms)
Bike						
TM	74.8 (1.1)	31.5 (2.1)	58.5 (3.3)	79.5 (2.3)	63.7 (5)	292.1 (20.4)
Basic Saliency	81.8 (2.5)	63 (4.3)	68.8 (6.2)	83 (5.5)	60.1 (3.8)	86.23 (12.2)
Biased Saliency	90.2 (2.4)	78.1 (3.1)	70.9 (4.2)	86.1 (5)	77.6 (3)	98.4 (10.5)
Crossing						
TM	75.5 (1.6)	37.9 (2.2)	64.2 (2.7)	72.4 (3)	75 (1.9)	319.8 (12.1)
Basic Saliency	78.2 (1.8)	39.2 (3.4)	59.2 (4.9)	80.8 (4.3)	86.4 (3.3)	92.66 (10.1)
Biased Saliency	95.8 (2.7)	48.2 (3.4)	62.2 (4)	88.3 (5.6)	93.1 (3.2)	102.7 (15.4)
Pedestrian						
TM	75.2 (2.9)	25.5 (2.7)	47.1 (4.1)	75 (3.9)	75.5 (4.3)	315.5 (14.3)
Basic Saliency	83.3 (1.5)	51.7 (4)	59.4 (3.9)	71 (6.1)	74.5 (5.9)	90.27 (12.5)
Biased Saliency	94.2 (1.5)	77.1 (2.5)	70.9 (3.5)	85 (3.5)	75.7 (5.3)	106.9 (14.2)
Average						
TM	75.1 (0.3)	31.6 (6.2)	56.6 (8.7)	75.6 (3.5)	71.4 (6.6)	309.1 (14.9)
Basic Saliency	81.1 (2.6)	51.3 (11.9)	62.4 (5.4)	78.2 (6.3)	73.6 (13.1)	89.7 (3.2)
Biased Saliency	93.4 (2.8)	67.8 (16.9)	68 (5)	86.4 (1.6)	82.1 (9.5)	102 (4.2)

Computation time is for detection of targets without noise. Numbers in parentheses are standard deviations

For example when asking a person to look for a clock in a room, he will probably search on walls first instead of ceil. In this section, we propose a simple heuristic to reduce the saliency computation of the biased saliency model by using historical knowledge of the agent.

Experiment IV. Offline learning of task-relevance saliency map (TSM) Road traffic signs are typically placed either by the roadside or above roads. They provide important information for guiding, warning, or regulating the behaviors of drivers in order to make driving safer and easier. In this experiment, we used the priori assumptions on image formation like when signs are photographed from a driver's position or assuming that road is approximately straight. This leads to ignoring large portions of the image when looking for signs. Motivated by these restrictions, we build a top-down task relevance map to consider such selectivity in space. Three salient locations were generated using biased saliency model for all training images for each traffic sign. A Gaussian mask was applied on each salient location to weight center locations more than surrounds. Then all maps were normalized and summed to form the final task-relevant spatial map shown in Fig. 11.

An advantage of such offline top-down task-relevance saliency map is that saliency computation and object detection could be started first from spatial areas, which have higher probability to contain an object and then to other areas. In order to use TSM for sign detection, saliency computation was limited in rectangular areas of Fig. 11. Over test sets of all three traffic signs, we were able to achieve the detection rates as before (Table 2), but about 3 times faster.

5 Discussions

The focus of this paper is on the learning phase of visual attention, where relevant feature values for a target are learned using several training images. Here the system automatically determines, which features to attend in order to best separate a target from its surroundings.

An issue in revised saliency model is setting window size (n). In our experiments in this paper, we set values for this parameter experimentally based on extent of an object in scenes. We are looking for systematic determination of it. In general any fast surround inhibition operation, which acts on a single scale can be used.

Our results show that color information has higher importance and information for discrimination of objects. Orientation features, which have rough information about structure of objects are in the second rank. Intensity does not seem to have high information for object detection. Overall, it is hard to judge, which scales are important in saliency detection from our results. However, it seems that it depends on the size of the target object relative to the image size. With small number of training images, our biased system was able to detect objects in large number of unseen test images. As Table 2 shows our system has higher detection rates than the basic saliency model in without-cost case. In with-cost case

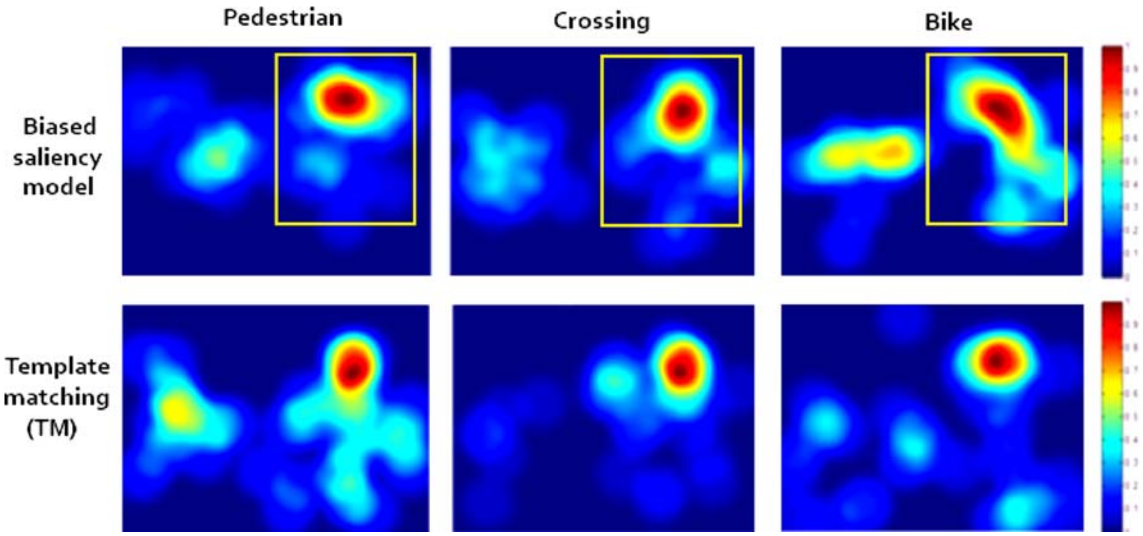

Fig. 11 Task-relevant saliency map (TSM). *Top row* shows offline learned biased saliency map averaged for all images in the train set for pedestrian, crossing and bike signs from *left* to *right*. Three FOA's were generated for building this map. *Bottom row* shows the same but with salient locations computed with template matching method. As could be seen both maps seem to have high correlation but making such a map with template matching takes about three times more computation. *Rectangles* show the areas that saliencies were computed for detection of a sign

Table 5 Average hit numbers of our biasing approach versus two other biasing approaches

	Target	# Train images	# Train images in VOCUS	# Test images	Avg. hit number		
					Our method	NVT [20]	VOCUS [19]
Campus map		9	2	7	1	1.2	1
Fire hydrant		8	2	8	1	1	1
Coke		45	5	59	1.6	3.8	1.3

while our system has higher detection rates compared to the basic saliency model it has lower cost. Over noisy images, our model performed better than template matching and basic saliency model. Its computation time is lower than that of template matching and slightly above the basic saliency model.

We also compared our attention system with two recent computational attention systems known as Neuromorphic Vision Toolkit (NVT) [20] and VOCUS [19]. These systems were chosen to be compared with because they are the most similar models to ours and are both based on the basic saliency model. Average hit numbers of our method and these approaches are shown in Table 5 in terms of the parameters that has been reported in [19].

As Table 5 shows our biasing method performed the same as VOCUS and NVT for fire hydrant detection. It was the same as VOCUS but better than NVT for campus map detection. It was better than NVT and slightly worse than VOCUS in detection of coke object. While sizes of train and test sets are small for the first two objects, it cannot be conclusive to judge, which method is better than others from these two objects. However, since sizes of train and test sets are larger for the third object, it proves that our method outperforms NVT and competes with VOCUS. Performances reported in Table 5 are derived in without-cost of our algorithm. While performance of our approach is slightly lower than VOCUS, it proposes a framework to incorporate cost of features in object detection. For example for an agent, it might be acceptable to tolerate a lower detection rate but with smaller computation time while performing a real time task. In other words, our approach enables an agent with certain computational resources to gain maximum detection rate.

6 Conclusions and future works

In this work, we introduced a new method for learning top-down attentional control from offline training data. Our approach detects targets robustly and quickly in various scenes. It is built over the basic saliency-based model of visual attention and biases it for synthetic saliency detection as well as natural object and traffic sign detection in cluttered scenes. It provides a method for quickly localizing object candidates to which computationally expensive recognition algorithms may be applied. Therefore, it provides a basis for robust and fast object recognition in computer vision and robotics. Since all feature channels and scales of the basic saliency model are not necessary for detection of an object, we put costs on computation of features of the model. The result was that, those channels which were not necessary and had not much effect on detection were not selected in the evolutionary process. Performance of our method was compared against basic saliency model as well as the template matching approach over noisy images.

A heuristic approach was also proposed for limiting the saliency computation over the most probable locations containing an object instead of the entire scene. For this, saliency maps from the previous experiences of the agent were added to form a top-down task relevance map. It was shown that this map reduces the computation time for traffic sign detection while having the same detection rate.

For our future research, we intend to learn top-down manipulations over the basic saliency model interactively and online using a reward and punishment mechanism. That way agent learns which objects to attend in each situation and then renders the image based on the top-down signals learned offline for that object. It would be also interesting to integrate the system into a robot control architecture enabling the detection of salient regions and goal-directed search in dynamic environments. In addition to the weights, there are also other parameters in the model, which can be tuned, for example number of feature channels, number of scales in the image pyramid, window size of the surround inhibition operation and number of color and orientation channels. Making advanced use of object structure for biasing or modifying the basic saliency model can also be an interesting future work.

References

1. Desimone, R., Duncan, J.: Neural mechanisms of selective visual attention. Ann. Rev. Neurosci. **18**, 193–222 (1955)
2. Li, Z.: A saliency map in primary visual cortex. Trends Cogn. Sci. **6**, 9–16 (2002)
3. Corbetta, M., Shulman, G.L.: Control of goal-directed and stimulus-driven attention in the brain. Nat. Rev. **3**, 201–215 (2002)
4. Posner, M.I.: Orienting of attention. Q. J. Exp. Psychol. **32**, 3–25 (1980)
5. Posner, M.I., Cohen, Y.: Components of visual orienting. Attention and Performance X, In: Bouma, H., Hillsdale, B.D. (eds.) Erlbaum, 531–556 (1984)
6. Maunsell, J.H., Treue, S.: Feature-based attention in visual cortex, Trends in Neurosciences. Neural Substr. Cogn. **29**, 317–322 (2006)
7. Kanwisher, N., Driver, J.: Objects, attributes, and visual attention: which, what, and where. Curr. Dir. Psychol. Sci. **1**, 26–31 (1992)
8. Yarbus, A.L.: Eye movements during perception of complex objects. In: Riggs, L.A. (ed.) Eye Movements and Vision, Chap. VII, pp. 171–196. Plenum Press, New York (1967)
9. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**, 1254–1259 (1998)
10. Nakayama, K., Maljkovic, V., Kristjansson, A.: Short term memory for the rapid deployment of visual attention. In: Gazzaniga, M.S. (ed.) The Cognitive Neurosciences, 3rd edn. MIT Press, Cambridge (2004)
11. Nakayama, K., Mackeben, M.: Sustained and transient components of focal visual attention. Vis. Res. **29**, 1631–1647 (1989)
12. Kristjansson, A., Nakayama, K.: A primitive memory system for the deployment of transient attention. Percept. Psychophys. **65**, 711–724 (2003)
13. Maljkovic, V., Nakayama, K.: Priming of popout I. role of features. Mem. Cogn. **22**, 657–672 (1994)
14. Rybak, I.A., Gusakova, V.I., Golovan, A.V., Podladchikova, L.N., Shevtsova, N.A.: A model of attention-guided visual perception and recognition. Vis. Res. **38**, 2387–2400 (1998)
15. Klein, R.M.: Inhibition of return. Trends Cogn. Sci. **4**, 138–147 (2000)
16. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Hum. Neurobiol. **4**, 219–227 (1985)
17. Peters, R.J., Itti, L.: Applying computational tools to predict gaze direction in interactive visual environments. ACM Trans. Appl. Percept. **5**(2), Article 8 (2008)
18. Navalpakkam, V., Itti, L.: Modeling the influence of task on attention. Vis. Res. **45**, 205–231 (2005)
19. Frintrop, S.: VOCUS: a visual attention system for object detection and goal-directed search, vol. 3899. PhD thesis, Lecture Notes in Artificial Intelligence (LNAI) (2006)
20. Navalpakkam, V., Itti, L.: An integrated model of top-down and bottom-up attention for optimizing detection speed. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 2049–2056 (2006)
21. Gavrila, D.M.: Traffic sign recognition revisited. In: Mustererkennung (DAGM). Springer, Bonn (1999)
22. Barnes, N., Zelinsky, A.: Real-time radial symmetry for speed sign detection. In: IEEE Intelligent Vehicles Symposium (IV), pp. 566–571, Parma, Italy (2004)
23. de la Escalera, A., Armingol, J.M., Mata, M.: Traffic sign recognition and analysis for intelligent vehicles. Image Vis. Comput. **21**, 247–258 (2003)
24. Paclik, P., Novovicova, J., Somol, P., Pudil, P.: Road sign classification using Laplace kernel classifier. Pattern Recognit. Lett. **21** (13–14), 1165–1173 (2000)
25. Gomez, L.C., Fuentes, O.: Color-based road sign detection and tracking. In: Proceedings Image Analysis and Recognition (ICIAR), Montreal (2007)
26. Fang, C.Y., Chen, S.W., Fuh, C.S.: Road-sign detection and tracking. IEEE Trans. Vehicular Technol. **52**(5), 1329–1341 (2003)
27. de la Escalera, A., Moreno, L.: Road traffic sign detection and classification. IEEE Trans. Ind. Electron. **44**, 848–859 (1997)

28. Burt, P.J., Adelson, E.H.: The Laplacian pyramid as a compact image code. IEEE Trans. Commun. **31**(4), 532–540 (1983)
29. Itti, L., Koch, C.: Feature combination strategies for saliency-based visual attention systems. J. Electron. Imaging **10**(1), 161–169 (2001)
30. Saliency toolbox homepage. http://www.saliencytoolbox.net/
31. Liang, J.J., Qin, A.K., Suganthan, P.N., Baskar, S.: Comprehensive learning particle swarm optimizer for global optimization of multimodal functions. IEEE Trans. Evol. Comput. **10**(3), 281–295 (2006)
32. Wolfe, J.M.: Visual search. In: Pashler, H. (ed.) Attention, East Sussex. Psychology Press, UK (1998)