ELSEVIER

Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE @ DIRECT®

Artificial Intelligence 146 (2003) 77–123

Artificial Intelligence

www.elsevier.com/locate/artint

Object-based visual attention for computer vision

Yaoru Sun [*], Robert Fisher

School of Informatics, The University of Edinburgh, 5 Forrest Hill, Edinburgh EH1 2QL, UK

Received 23 April 2002

Abstract

In this paper, a novel model of object-based visual attention extending Duncan's Integrated Competition Hypothesis [Phil. Trans. R. Soc. London B 353 (1998) 1307–1317] is presented. In contrast to the attention mechanisms used in most previous machine vision systems which drive attention based on the spatial location hypothesis, the mechanisms which direct visual attention in our system are object-driven as well as feature-driven. The competition to gain visual attention occurs not only within an object but also between objects. For this purpose, two new mechanisms in the proposed model are described and analyzed in detail. The first mechanism computes the visual salience of objects and groupings; the second one implements the hierarchical selectivity of attentional shifts. The results of the new approach on synthetic and natural images are reported.

Keywords: Visual attention; Object-based visual attention; Integrated competition; Grouping salience; Hierarchical selectivity

1. Introduction

It is well known that the primate visual system employs an attention mechanism to limit processing to important information that is currently relevant to behaviours or visual tasks. It can efficiently deal with the balance between computing resources, time cost and performing different visual tasks in a normal, cluttered and dynamic environment [68]. Visual attention selectivity can be either overt to drive and guide eye movements, picking up useful information over time [31,52,61], or covert, internally shifting the focus of attention from one locus to another without eye movements [44,67, pp. 519–570].

[*] Corresponding author.
E-mail address: yaorus@dai.ed.ac.uk (Y. Sun).

0004-3702/03/$ – see front matter © 2003 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0004-3702(02)00399-5

1.1. General problems of modelling visual attention

Modelling visual attention is a challenging problem for machine vision. Three closely-related basic questions are immediately identifiable:

(1) How can the visual system know what information is important enough to capture attention?
Modern research on visual attention from psychophysical and neurophysiological experiments have found that there exist two ways by which information can be used to direct attention (see [92,94] for reviews). One approach uses bottom-up information including basic features such as colour, orientation, motion, depth, conjunctions of features such as objects in 2D or 3D space and even learned features. In this case, visually salient features (a feature or stimulus differs from its immediate surround in some dimensions and the surround is reasonably homogeneous in those dimensions [19]) are mostly used to attract visual attention. A great number of models make use of "saliency" to direct attention [1,2,8,53,86]. However, saliency cannot always capture attention in a purely bottom-up fashion if attention is focused or directed elsewhere in advance [92,94]. Thus it is necessary to recognize the importance of how attention is also controlled by top-down information relevant to current visual behaviours. The deployment of attention is determined by an interaction between bottom-up input and top-down attentional priming or setting [94].

(2) How does the visual system know when and how to direct attention and choose important information rather than doing so at random times and by random selection?
This is the paradox of intelligent selection of attention in visual systems. We would like to know whether selection happens earlier or later, to what extent visual processing is serial or parallel, and what interplay exists between these factors. A number of researchers have proposed two-stage models in which the preattentive stage performs independent detection or extraction of primary visual features automatically in parallel (without attention) and the second stage of attention processes the combination of primitive features by serially shifting the focus of attention to scan subsets of the incoming information available from the previous stage (see [92] for a review). This proposed strategy, however, conflicts with many modern psychophysical experiments that confirm that attention can arise from very early visual processing stages (e.g., feature detection) or arise from relatively late processing stages (e.g., object representation or recognition) in different circumstances in which parallel and serial processing reciprocally intertwine for efficient performance of visual tasks [43,57,60]. Thus, this problem is far from well understood and requires further investigation.

(3) Where is (are) the next potential target(s) of visual attention shifts? That is, how does attention know where to go and what to do next?
There are two traditional assumptions in the literature attempting to account for this. The space-based attention theory holds that attention is allocated to a region of space, with processing of everything within this spatial window of attention like a spotlight, internal eye, or zoom-lens [28,70,83,84]. Object-based attention theory argues that attention is actually directed to an object or a group of objects to process any properties of selected object(s) rather than regions of space [16,18,48,76]. Some recent findings

support a view that the two accounts are not mutually exclusive [22,30,43] and they may actually share common neural mechanisms in the parietal lobes [32]. Until now, few researchers have proposed attentional models that integrate space-based and object-based views (but see [59]). As suggested by S.E. Palmer in [67, pp. 547–549], both hypotheses may be true to account for different processing levels respectively in the visual system and may be necessary to supply and interact at multiple processing levels for coherent behaviour.

These three problems lead to a general question: How does visual attention work to perform efficient selectivity? The dominant theory of visual attention is based on the hypothesis that attention works in space like a "spotlight" or "zoom lens", scanning the scene by shifting attention from one location to the next to limit processing to a variable size of space in the visual field. There have been a number of attentional models that use this hypothesis. Most of them are derived from Treisman's Feature Integrated theory [83] which consists of separate low-level feature maps that are combined together by a spatial attention window operating on a master map or saliency map. We will briefly review the most influential accounts of visual attention in psychophysics and the correspondingly inspired computable models below.

1.2. Psychophysical models of attention

There are two divisions of theories in the vast literature being developed to understand visual attention. One is the very influential space-based attention theory. Another is the developing theory concerning object-based attention. So far Treisman's model is the most successful model of space-based attention and does provide a general framework for understanding visual attention. Following her theory, a number of computational models of attention in the psychophysics and computer vision fields have been developed. The main difference between them is that they use different methods to construct and combine the low-level feature maps and to model the control mechanisms of attentional movements. In addition, there are lots of other well-known models of spatial attention as well, such as the guided search model of Wolfe [91], the spotlight or zoom lens model of Eriksen et al. [27,28], the saliency map model of Koch and Ullman [53], the dynamic routing model of Olshausen et al. [66] and the like.

The essential bifurcation between object-based attention and space-based attention lies in the question of what are the underlying units of attentional selection. In contrast to the traditional models of space-based attention, object-based attention holds that visual attention can directly select discrete objects rather than only and always selecting continuous spatial areas of the visual field. The research on object-based attention is still quite new. However, some fundamental theories have been developed in recent years. In the "Biased Competition Model" of Desimone and Duncan [14] and the "Integrated Competition" hypothesis of Duncan [23], visual attention is taken as an emergent effect of competition between neural representations in multiple systems which work together to serve the same selected object. Other pioneer research can be seen in the work of Humphreys and his colleagues [19,41,42], Grossberg [39], Behrmann [4], and a converged review [76].

1.3. Computable models of space-based attention

Koch and Itti have built the most sophisticated saliency-based spatial attention model [45,53]. The saliency map is used to encode and combine information about each salient or conspicuous point (or location) in an image or a scene to evaluate how different a given location is from its surrounding. A Winner-Take-All (WTA) neural network implements the selection process based on the saliency map to govern the shifts of visual attention. This model performs well on many natural scenes and has received some support from recent electrophysiological evidence [34,74]. Tsotsos et al. [86] presented a selective tuning model of visual attention that used inhibition of irrelevant connections in a visual pyramid to realize spatial selection and a top-down WTA operation to perform attentional selection. In the model proposed by Clark et al. [8,9], each task-specific feature detector is associated with a weight to signify the relative importance of the particular feature to the task and WTA operates on the saliency map to drive spatial attention (as well as the triggering of saccades). In [36,73], colour and stereo are used to filter images for attention focus candidates and to perform figure/ground separation. Grossberg proposed a new ART model for solving the attention-preattention (attention-perceptual grouping) interface and stability-plasticity dilemma problems [37,38]. He also suggested that both bottom-up and top-down pathways contain adaptive weights that may be modified by experience. This approach has been used in a sequence of models created by Grossberg and his colleagues (see [7,38] for an overview). In fact, the ART Matching Rules suggested in his model tends to produce later selection of attention and is partly similar to Duncan's integrated competition hypothesis [22] which is an object-based attention theory and different to the above models.

Some researchers have exploited neural network approaches to model selective attention. In [2,3], the saliency maps which are derived from the residual error between the actual input and the expected input are used to create the task-specific expectations for guiding the focus of attention. Kazanovich and Borisyuk proposed a neural network of phase oscillators with a central oscillator (CO) as a global source of synchronization and a group of peripheral oscillators (PO) for modelling visual attention [51]. Similar ideas have also been found in other work [11,12,55,63,64] and are supported by many biological investigations [55,79,87]. There are also some models of selective attention based on the mechanisms of gating or dynamic routing information flow by dynamically modifying the connection strengths of neural networks [37,42,66,71].

In some models, mechanisms for reducing the high computational burden of selective attention have been proposed based on space-variant data structures or multiresolution pyramid representations and have been embedded within foveation systems for robot vision [6,10,29,75,80,82,90]. But it is noted that these models developed the overt attention systems to guide fixations of saccadic eye movements and partly or completely ignored the covert attention mechanisms. Fisher and Grove [40] have also developed an attention model for a foveated iconic machine visual system based on an interest map. The low-level features are extracted from the currently foveated region and top-down priming information are derived from previous matching results to compute the salience of the candidate foveate points. A suppression mechanism is then employed to prevent constantly re-foveating the same region.

1.4. Inducements and innovations of the proposed model

The computable models of space-based attention reviewed above, however, have some intrinsic disadvantages. They have only concentrated on mechanisms of visual attention based on selectivity by spatial locations. Thus they inherently lack mechanisms accounting for object-based selection (see [24,25] and [67, pp. 547–549] for reviews). The normal scene is usually cluttered: objects may overlap or share some common properties. In this case attention may need to work in several discontinuous spatial regions at the same time. Some different visual features which constitute the same object can come from the same region of space: in this case maybe no attention shift is required. The structure of one object may be very complex and hierarchical: in this case the interaction or cooperation of object-based, location-based selection and selectivity by visual features is required. Object-based attention has advantages that space-based attention does not have:

- more efficient visual search: speed and accuracy;
- less chance to select a nonsense or empty location;
- naturally hierarchical selectivity.

Thus it is important to properly integrate the two accounts of space-based and object-based attention.

The above problems led us to propose another machine vision approach for modelling visual (covert) attention. The model described here is an alternative computational model of visual attention which is object-based. It absorbs several ideas and many findings from modern literature in psychophysics and computer vision, including recent research on: (1) object-based visual attention such as Duncan's Integrated Competition theory [21–23], and [14,15]; (2) visual saliency such as Koch and Itti's model of saliency-based visual attention [45,53]; (3) bottom-up and top-down interaction of visual attention [92,94]; (4) integration of object-based and location-based attention [59]; (5) visual representations of within-objects and between-objects [43]; and (6) other investigations [5,38].

One of the novel mechanisms in our model is the grouping-based salience computation for attentional competition between features, objects, and groupings of features and objects, and competition within objects and groupings of features and objects. The early visual features of the scene (colours, intensity, and orientations) are extracted by multiresolution pyramids. The visual salience of points, objects and regions is calculated for different groupings on the feature pyramids, which builds up the basis of the purely bottom-up attention competition among various visual inputs. The competition for visual attention is modulated by the interaction between bottom-up visual saliency and the top-down attentional setting which is decomposed into positive priming, negative priming, free, and occupied cases (introduced later). *The main goal of this paper is to present our model for the visual saliency of groupings and the mechanism of covert attentional movements*.

Another novel mechanism used in the proposed model is hierarchical selectivity for guiding covert attentional movements, which can be regarded as a kind of multiple selectivity [67, pp. 547–554] integrating attentional selection by spatial locations, visual features and their complex conjunctions (e.g., objects or groupings). The competition for attention takes place firstly from the most coarse level on multiresolution pyramids, then gradually to the finer level, as well as from coarser groupings to finer groupings within and between groupings and resolutions. The finest grouping is set to a pixel or point in our model. This mechanism is thus biologically plausible. Clearly, the strategies from coarse to fine occur on the multiple architecture of visual resolution and groupings including objects, features and locations related to the relevant resolution. At each pyramid level, the winner of selective attention in each competition is generated by a Winner-Take-All (WTA) strategy.

The presented model explores the first machine-vision implementation of a hierarchical object-based visual attention system. The paper shows that it produces plausible attention shifts on real imagery and also that its performance on synthetic displays is similar to human psychophysical results. To simplify the research, we have assumed that a perceptual organisation of the image into a hierarchical set of groupings has been done. (We assume that other research from elsewhere will eventually supply this input. See Section 2.6 for the further discussion.) Further, our approach has a mechanism to respond to top-down behavioural inputs, but we have not completely investigated the actual top-down selection process (as this is a complex process involving both visual and non-visual reasoning). Lastly, the presented model only considers covert attention (where the fovea does not move) rather than overt eye movements that might lead to significant changes in visual salience.

2. Model

2.1. Overview of the model

Our work is concerned with the development of efficient mechanisms of visual attention for a machine vision system. The model developed here shows that object-based and location-based attention can work in a uniform framework depending on both the current scene and the observer's goals to deal with complex visual tasks (see [76] for a comprehensive review of object-based visual attention). The model, for this purpose, brings together several issues found in the modern literature. The critical aspects of our theory are:

(1) *Integrated competition for visual attention.*

Our approach extends Duncan's Integrated Competition hypothesis [15,22,23]. The main adjustment is that we think his model of object-based attention can be extended to work in both object-based and space-based fields by replacing object-centered with grouping-centered (see one of the few psychophysical attentional models [5] and [38] for integrating object-based with location-based evidence). A grouping is a unit involving object(s) and related features and locations (see [17,43,76] for detailed discussion on these issues). In this way, a grouping in our model can be a point (a pixel here), an object or a feature, a group of objects or features, or a region. At any given moment, enhanced responses to one grouping will decrease responses to other competitors. Once one grouping gains the dominance of selective attention, all other relevant processing to this grouping and all components belonging to this grouping share the same dominance. This is why it is termed “integrated competition”.

(2) *Bottom-up and top-down interaction.*

The nature of attentional competition comes from dynamic interaction based on visual saliency [45,53] between bottom-up visual grouping and top-down attentional biasing or setting [92,94]. That is, purely bottom-up or top-down driven information for attention can only bias the competition for selection process partly. In this case, salient visual groupings can capture attention quickly and automatically only if the current attention is not deliberately directed to other groupings or properties in advance.

(3) *Hierarchical selectivity of visual attention.*

Hierarchical selectivity is proposed to guide the attentional movements shifting from one locus of attention to another under multiscale transformation and directly builds upon the above two issues. It implies that visual attention can directly select a continuous area of space, discrete object(s), feature(s), point(s), or their grouping. The space-based and object-based attentional selectivity are either cooperative or independent of each other for efficient selective acts according to the current visual situations and tasks. This strategy is especially useful for machine vision. For example, space-based selection can be applied to region segmentation whereas object-based selection can be used for object recognition or fine analysis.

Keeping the above issues in mind, our model of visual (covert) attention is depicted schematically in Fig. 1. The model firstly extracts primary features colours, intensity, and orientations from one fixated image sampled from a given scene, by multiscale pyramid filters. After perceptual grouping preprocessing, the bottom-up saliency mappings of various groupings are created via the grouping-based salience computation. These saliency mappings are dynamically varied according to competition conditions among the groupings at different resolutions and related surroundings during the attentional movements. The results produced from this stage are fed to the attention competition pool where all coarsest groupings compete against each other for preferentially obtaining the selective attention. The competition procedure is a dynamic interaction between bottom-up salience and the top-down attentional setting. The rules of winner-take-all and inhibition of return are applied here to ensure the winner benefits and prevent attention from returning to the previously attended groupings. The attentional movements among the winning competitors are guided by hierarchical selectivity. The detailed description of each module in the model is given in the following sections.

2.2. Eye/fixation image

Our model is built for covert visual attention rather than overt eye movements such as gaze control in active vision research. At any moment, a fixed image, which is a transformation of the world image into retinal image at each fixation point, is obtained by simulating the functional mapping of resolution decreasing from the fovea to the periphery

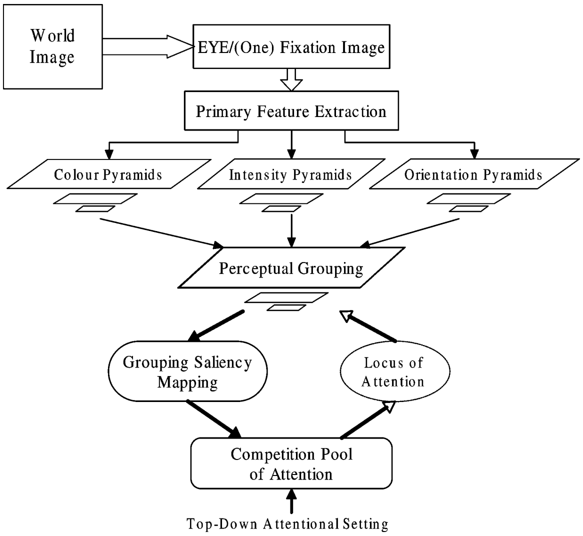

Fig. 1. The schematic description of the model.

of the retina. The following modules involved in visual attention operate on one given fixed image and the sampling function of a gaze at that moment is presented as follows. Future research will consider overt eye movements such as saccades and how saliency is integrated with overt multiple fixations.

2.3. Primary feature extraction

Colour image input is decomposed into sets of multiscale feature maps via overcomplete steerable pyramid filters [35], to generate four colour, one intensity, and four orientation pyramids [45]. Suppose that F is the input image, with r, g, b being the red, green, and blue colour components of F. An intensity image $I(p_{ij})$ is created by:

$$I(p_{ij}) = \big[r(p_{ij}) + g(p_{ij}) + b(p_{ij})\big]/3 \tag{1}$$

where p_{ij} is a point of F, $i \in [1\ldots n]$, $j \in [1\ldots m]$, $n \times m$ is the size of the image.

Then, four colour channels R (red), G (green), B (blue), and Y (yellow) are obtained as [45] (negative values are set to zero):

$$\begin{aligned}
R(p_{ij}) &= r(p_{ij}) - \big[g(p_{ij}) + b(p_{ij})\big]/2,\\
G(p_{ij}) &= g(p_{ij}) - \big[r(p_{ij}) + b(p_{ij})\big]/2,\\
B(p_{ij}) &= b(p_{ij}) - \big[r(p_{ij}) + g(p_{ij})\big]/2,\\
Y(p_{ij}) &= \big[r(p_{ij}) + g(p_{ij})\big]/2 - \big|r(p_{ij}) - g(p_{ij})\big|/2 - b(p_{ij}).
\end{aligned} \tag{2}$$

Let W_{lpf}, $W_{bpf}(\lambda;\theta)$ be Gaussian and orientated Gabor steerable filters respectively. With these filters acting on the five $I, R, G, B,$ and Y channels (see [35,45] for more details), we can construct intensity, colour (red, green, blue, and yellow) and orientation pyramids:

$$I_{\lambda+1} = W_{lpf}^T \cdot W_{lpf} \cdot I_\lambda;\ I_0 = I, \tag{3}$$

$$R_{\lambda+1} = W_{lpf}^T \cdot W_{lpf} \cdot R_\lambda;\ R_0 = R,$$

$$G_{\lambda+1} = W_{lpf}^T \cdot W_{lpf} \cdot G_\lambda;\ G_0 = G,$$

$$B_{\lambda+1} = W_{lpf}^T \cdot W_{lpf} \cdot B_\lambda;\ B_0 = B,$$

$$Y_{\lambda+1} = W_{lpf}^T \cdot W_{lpf} \cdot Y_\lambda;\ Y_0 = Y, \tag{4}$$

$$O_\lambda(\theta) = W_{bpf}(\lambda; \theta) \cdot I, \tag{5}$$

where $\lambda \in [1 \ldots l]$ is the pyramid's scale and $\theta \in [0°, 45°, 90°, 135°]$ or $[0°, 22.5°, 45°, 67.5°, 90°, 112.5°, 135°, 157.5°]$ (in this paper, we used both orientation sets for different experiment environments but the first is the general one) is the preferred orientation. The Anderson kernel used for W_{lpf} is $(1/16, 1/4, 3/8, 1/4, 1/16)$. The Gabor filter comes from modulating the related Lapacian pyramids with a set of oriented sine waves, then being followed by low-pass operation, and finally taking the modulus (see [35] for these two filters in detail).

2.4. Grouping-based saliency mapping

Salience evaluation based on groupings is the bridge to achieve object-based attention and integrate space-based attention in this paper. In our approach, groupings are the primary perceptual units upon which attentional processes operate. The term "grouping" (or "segmentation") is a common concept in the long research history of perceptual grouping by the Gestaltists (see [67, pp. 257–266] for a review). We evaluate salience based on groupings here because "grouping" itself has already embedded "object" and "space". This usage constitutes a fundamental difference to most of the previous computable models of space-based attention.

A grouping is a hierarchical structure of objects and space. In this sense, a grouping may be a point, an object, a region, or a hierarchical structure of groupings. However, we are not implementing the grouping process in this paper but assume in the work below that it exists. That is, we assume that a given scene at each scale has already been segmented into groupings according to the Gestalt principles (or other grouping approaches). Some further discussions on grouping are given in Section 2.6. The theory proposed here for salience computation is independent of the approach used for perceptual grouping.

The salience of a grouping is a function of all saliency contributions coming from the components within the grouping working together to compete with their common competitors and competing with each other. This notion covers two issues. One issue is the relationship of spatial location, objects, and features to the grouping they belong to, as shown in Fig. 2. The figure shows that grouping salience is computed from its components of spatial location, feature(s), and/or object(s). The other issue is the competition between a grouping and its surroundings by cooperation and the competition between its components. The effect of a competition between two competitors may either enhance or suppress their salience according to their contrast properties (Fig. 3(a)). Two simple examples are given in Figs. 3(b) and (c).

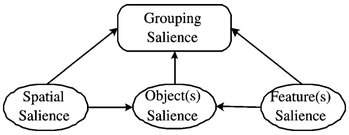

Fig. 2. Diagram of grouping salience.

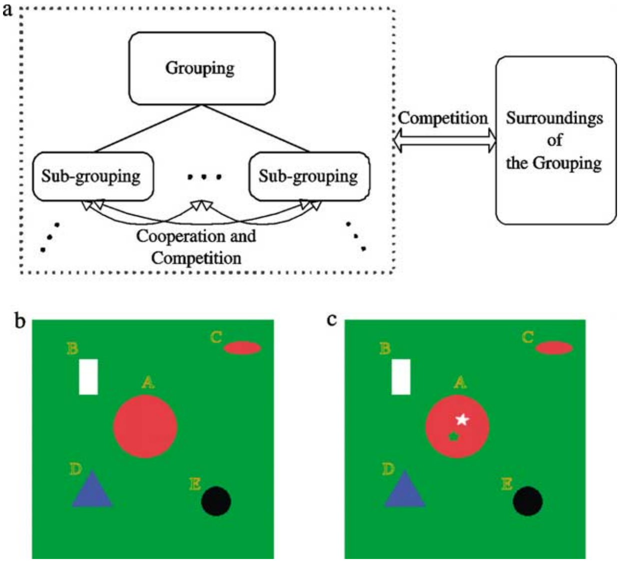

Fig. 3. An example of grouping salience.

Suppose the red circle (grouping A) is the target. We want to calculate its salience. Its surroundings consist of four groupings (B, C, D, E) and other background points (green pixels). In Fig. 3(b), all red pixels within grouping A work together to enhance the grouping salience by feature contrast to compete with the surroundings. Along with this global competition, local competitions among pixels within grouping A also produce a negatively enhanced effect on the grouping salience due to the same features of these pixels. In Fig. 3(c), the green star sub-grouping in grouping A brings a suppressive effect on the total A grouping salience when it competes with green pixels in the background but a enhanced effect when it competes with the non-green groupings and pixels with A and elsewhere. The final salience of grouping A depends on the competitive effects brought by all of the components within A (including red pixels, white star and green star).

Based on the above considerations, the contrast between any two points is the primitive operation in the computation of grouping salience. However, we are not claiming that the salience computation theory introduced below is complete. The paper in fact is concerned with salience deriving from the colour, intensity, orientation, and distance factors only. Many other factors affecting salience are not included here, such as motion, shape, size, depth and the like (see [92] for the related issues in visual search). One unconsidered factor about relative size difference between groupings will be discussed later (see Section 2.4.3).

The salience of a grouping is calculated by combining the colour, intensity, and orientation salience of the components of the grouping. Due to the close relationship between the chromatic opponent-colour channels and the achromatic (white-black) channel in the visual perception and contrast process [88,89], we calculate colour and intensity salience together.

Suppose $\Re$ is any given grouping at the current resolution scale λ at time t, Θ is the surroundings of $\Re$. If $\forall \Re_i \in \Re$, $\forall \Re_j \in (\Re \cup \Theta)$ and $i \neq j$, we calculate the colour-intensity salience S_{CI} and orientation salience S_O of $\Re_i$ by:

$$\begin{aligned} S_{CI}(\Re_i;\lambda;t) &= f_{CI}\big(\Re_i;\{\Re_j\};\lambda;t\big),\\ S_O(\Re_i;\lambda;t) &= f_O\big(\Re_i;\{\Re_j\};\lambda;t\big), \end{aligned} \tag{6}$$

where f_{CI}, f_O are the calculating function of colour-intensity, orientation salience between $\Re_i$ and $\Re_j$ respectively. The salience S of grouping $\Re$ is given as:

$$\begin{aligned} S(\Re_i;\lambda;t) &= \Gamma\big[S_{CI}(\Re_i;\lambda;t); S_O(\Re_i;\lambda;t)\big],\\ S(\Re;\lambda;t) &= \Psi\big[S(\Re_i;\lambda;t)\big], \end{aligned} \tag{7}$$

where Γ, Ψ are normalization and integration functions respectively. These functions are defined in detail below.

This computational model of saliency is built upon the principle of localization, relativity and the dynamics of visual input in the scene as covert attention occurs. As pointed out in [34], most (stable) objects in a normal environment are not intrinsically salient but can become salient if they are behaviourally significant. The normal scene has a hierarchical structure, thus features may not always have the same salience when viewed in extended regions or larger contexts. In other words, the salient difference among objects or features may change over time, or as background or the context of the scene changes. The saliency computation is a complex and difficult problem. Until now few research studies in the field of attention in machine vision have dealt with it (however, see [45,46,77] for some discussion related to spatial saliency map). From our point of view, visual saliency arises from the competition between different groupings and between a grouping and its surroundings.

For simplicity in formulas, all computations below are defined for a given current time and resolution scale. The salience computation at other times and spatial scales is similar because the salience of a grouping is decided only by the current constitution of the grouping and its surroundings. Thus the changing of salience over time (salience dynamics) of a grouping depends upon the varying of the grouping's current constitution and surroundings over time. That is, the same computation rules are used for any time and scale when the segmentation of groupings at that time and scale is given. In this way, the full details of the computable approach are given below.

2.4.1. Colour and intensity salience

Assume x, y are two arbitrary pixels in a grouping $\Re$ on level λ of pyramids of colours, intensity, and orientations. Then, the properties of x and y can be denoted by a tensor composed of a 4-dimension colour vector, a 1-dimension achromatic intensity vector, and a 4-dimension orientation vector. For example, pixel

$$\begin{aligned} x = \big(&\{R_{x,\lambda,\Re}, G_{x,\lambda,\Re}, B_{x,\lambda,\Re}, Y_{x,\lambda,\Re}\}, \{I_{x,\lambda,\Re}\}, \\ &\{O_{x,\lambda,\Re}(\theta_1), O_{x,\lambda,\Re}(\theta_2), O_{x,\lambda,\Re}(\theta_3), O_{x,\lambda,\Re}(\theta_4)\}\big). \end{aligned}$$

In the following section we suppose all calculations are within a given group on a given pyramid level, so the subscripts $\Re$ and λ will be generally omitted. We first compute the property contrast between pixels x and y. Let RG and BY be the two colour "double-opponent channels" of red-green/green-red and blue-yellow/yellow-blue [26,47], so we have:

$$\begin{aligned} RG(x, y) &= \big|(R_x - G_x) - (R_y - G_y)\big|/2, \\ BY(x, y) &= \big|(B_x - Y_x) - (B_y - Y_y)\big|/2. \end{aligned} \tag{8}$$

The colour chromatic contrast ΔC between x and y is calculated as:

$$\Delta C(x, y) = \sqrt{\eta_{RG}^2 RG^2(x, y) + \eta_{BY}^2 BY^2(x, y)} \tag{9}$$

where η_{RG} and η_{BY} are the weighting parameters. In this paper, we set them as:

$$\begin{aligned} \eta_{RG} &= \frac{R_x + R_y + G_x + G_y}{R_x + R_y + G_x + G_y + B_x + B_y + Y_x + Y_y}, \\ \eta_{BY} &= \frac{2\sqrt{B_x^2 + B_y^2 + Y_x^2 + Y_y^2}}{3 \times 255}, \end{aligned} \tag{10}$$

where the 255 parameter is used here because of the representations of colour and intensity in this paper have the maximum value 255. The weights η_{RG} and η_{BY} can be optimized further according to more colour discrimination experiments or references in the colour research literature. The results produced by setting η_{RG} and η_{BY} as those in formulae (10) are very close to L*u*v* (see [62,72] for related issues). We obtain equal maximal contrasts between opponent colours such as red and green, blue and yellow, or white and black. The contrasts between other colours are also reasonable. For example, it is acceptable that the colour contrast between yellow and black is greater than yellow and white, etc. (see [49,62,93] for more discussion). All values of colour-intensity contrasts between x and y fall into the range $[0 \ldots 255]$.

The intensity contrast between the two pixels x and y is:

$$\Delta I(x, y) = \big|I(x) - I(y)\big|. \tag{11}$$

So, the formula for calculating salience $S_{CI}(x, y)$ of colour-intensity between x and y is:

$$S_{CI}(x, y) = \sqrt{\alpha \Delta C(x, y)^2 + \beta \Delta I(x, y)^2}, \tag{12}$$

where α and β are weighting coefficients and we here set them to 1.

Suppose d_{gauss} is the Gaussian distance function between x and y. The Gaussian distance is defined as:

$$d_{\text{gauss}}(x, y) = \left(1 - \frac{\|x - y\|}{\hat{n} - 1}\right) \mathrm{e}^{-\frac{1}{2\sigma^2}\|x-y\|^2} \tag{13}$$

with the scale σ and distance $\|x - y\|$. In the experiments in this paper, the Gaussian scale σ is set to $\hat{n}/\rho$ where $\hat{n}$ is the maximum of the width and length of the feature maps on the current pyramid level λ. ρ is a positive integer and generally $1/\rho$ may be set to a percentage of $\hat{n}$, such as 2%, 4%, 5%, or 20%, 25%, 50%, etc. The greater ρ is, the smaller the radius between the neighborhood and its surrounding center is. In this way, the Gaussian distance guarantees competition throughout the attention window but the strength varies with distance. This function produces strong local competition between short-range neighbours and weak competition between long-range neighbours. Such similar effects of attention competition have been found in visual cortex [15]. Research on cortico-cortical connections shows that inhibition from the surround of the same stimulus properties as the center is strongest [78]. The distance $\|x - y\|$ can be the Euclidean distance but we prefer a chessboard distance: $\|x - y\| = MAX(|i - h|, |j - k|)$, (i, j), (h, k) are the coordinates of x, y on the current pyramid level. MAX denotes the maximizing operator. The reason for selecting the chessboard distance is that with the aid of this operator, the neighbours within the same, 8-adjacency neighbourhood have equal distance effects on their common center and the "center-surround" function can be easily simulated.

Let $\mathcal{NH}_{CI}$ be the neighbourhood surrounding x, $y_i \subset \mathcal{NH}_{CI}$ $(i = 1 \ldots n \times m - 1)$ be a neighbour. We use the following formula to calculate the colour-intensity salience of x:

$$S_{CI}(x) = \frac{\sum_{i=1}^{n\times m-1} S_{CI}(x, y_i) \cdot d_{\text{gauss}}(x, y_i)}{\sum_{i=1}^{n\times m-1} d_{\text{gauss}}(x, y_i)}. \tag{14}$$

2.4.2. Orientation salience

We define $\bar{\theta}_{x,y}$ as the orientation difference between pixels x and y. Let $u_x(\theta)$ and $u_y(\phi)$ be the orientation vectors of x and y in the current orientation pyramid respectively. Note that u, θ, and ϕ themselves all consist of multiple components. For example, $u_x(\theta) = [u_x(0), u_x(\pi/4), u_x(\pi/2), u_x(3\pi/4)]$, if we have four preferred orientations. We define the orientation salience $C_O(x, y)$ of x to y as:

$$C_O(x, y) = d_{\text{gauss}}(x, y)\sin(\bar{\theta}_{x,y}) \tag{15}$$

where d_{gauss} has already been defined in Eq. (13). A major reason that we select a sinusoid function for orientation contrast is that this function is a nonlinear and monotonically increasing function from 0 to 1 over the range $[0, \pi/2]$ and symmetric in $[0, \pi]$. Nothdurft has suggested that the salience of pop-out targets has a nonlinear (enhanced) character from threshold and saturation effects with increasing orientation contrast from 0 to $\pi/2$ [65]. If u_x and u_y have orientation strengths at all orientations, then the general calculation for $\bar{\theta}_{x,y}$ can be given by:

$$\bar{\theta}_{x,y} = \frac{\int_0^\pi \phi[\int_0^\pi u_x(\theta)u_y((\theta+\phi) \bmod \pi)\,\mathrm{d}\theta]\,\mathrm{d}\phi}{\iint_0^\pi u_x(\theta)u_y((\theta+\phi) \bmod \pi)\,\mathrm{d}\theta\,\mathrm{d}\phi}. \tag{16}$$

For practical computation in this paper, we give the following discrete form for $\bar{\theta}_{x,y}$:

$$\bar{\theta}_{x,y} = \frac{\sum_{j=0}^{\zeta-1} j\varphi \sum_{i=0}^{\zeta-1} u_x(i\varphi)u_y((i\varphi + j\varphi) \bmod \pi)}{\sum_{j=0}^{\zeta-1}\sum_{i=0}^{\zeta-1} u_x(i\varphi)u_y((i\varphi + j\varphi) \bmod \pi)} \tag{17}$$

where mod is the standard modulus operator, ζ is the number of orientation pyramids or preferred orientations, $\varphi = \pi/\zeta$. When ζ is 4 or 8, φ is $\pi/4$ or $\pi/8$.

The salience computation for orientation is more complicated than for colour-intensity. It is most important to take into account the homogeneity/heterogeneity of the neighbourhood of each point which is currently taken as a center for center-surround calculation. Psychophysical findings show that "pop-out" is closely related to the distribution of orientations in the local neighbourhood [56,68,83,85,92]. Aiming at a practical computation of orientation salience, further considerations of "center-surround" operations are provided as follows.

Let y_i ($i = 1 \ldots n_k$, n_k is the number of neighbours in the kth neighbourhood) be a neighbour in the distance k or kth neighbourhood $\mathcal{NH}_O(k)$ surrounding x. It is clear that the distance 1 or first neighbourhood of x has 8 closest neighbours surrounding x, and that the distance k neighbourhood has $8k$ neighbours. A boundary check must be applied to ensure all data comes from within the current image layer. Then the average orientation contrast of x to its kth neighbourhood is:

$$\overline{C}_O\bigl(x, \mathcal{NH}_O(k)\bigr) = \frac{1}{n_k} \sum_{y_i \in \mathcal{NH}_O(k)} C_O(x, y_i). \tag{18}$$

Suppose n_0 is the number of different directions within $\mathcal{NH}_O(k)$, we have $\omega_k = n_0 - 1$. This is used for checking and evaluating how heterogeneous the orientations are in the neighbourhood of x. n_0 can be obtained by a simple method: set $n_0 = 0$; then $n_0 = n_0 + 1$ if the orientation on which y_i has the maximum value on all orientation maps (this means the maximum sub-orientation vector of y_i is on that map) is different from the maximum sub-orientation vector of y_{i+1}.

We use the same set of histograms above to evaluate the orientation homogeneity of the whole surround of x. Let w_{ijk} be y_i's value on the orientation (θ_j) feature maps on kth neighbour "ring", n_r be the number of "rings" in the whole neighbourhood of x, then the method to calculate homogeneity weight ω for the whole surround is given in formulae (20).

Under these considerations, we have the orientation contrast of x to its kth neighbourhood:

$$\widehat{C}_O\bigl(x, \mathcal{NH}_O(k)\bigr) = \frac{\overline{C}_O(x, \mathcal{NH}_O(k))}{\xi + \omega_k}, \tag{19}$$

where ξ is a parameter used to prevent a zero denominator and usually set to 1.

Let m_r be the number of "rings" in a neighbourhood, and $d_{\text{gauss}}(k)$ (defined in Eq. (13)) be the Gaussian distance of the kth neighbourhood to x. Because of the chessboard distance, $d_{\text{gauss}}(k)$ is the same for each point within x's kth neighbourhood. Finally, the orientation salience of x to all of its neighbours is:

$$C_O(x) = \frac{\sum_k \widehat{C}_O(x, \mathcal{NH}_O(k)) \cdot d_{\text{gauss}}(k)}{(\xi + \omega) \cdot m_r \cdot \sum_k d_{\text{gauss}}(k)}$$

where

$$m_r = \sum_k 1 \quad \text{and} \quad \bigl|\widehat{C}_O\bigl(x, \mathcal{NH}_O(k)\bigr)\bigr| > 0;$$

ω is given by:

$$\omega = \sum_j \widehat{H}(\theta_j); \quad \widehat{H}(\theta) = \{\widehat{H}(\theta_j)\} = \left\{ \sum_k \frac{|H_k(\theta_j) - \overline{H}(\theta_j)|}{MAX\{H_k(\theta_j), \overline{H}(\theta_j)\}} \right\};$$

$$H_k(\theta_j) = \sum_{y_i \in \mathcal{N}\mathcal{H}_O(k)} w_{ijk}(\theta_j, y_i); \quad \theta_j \in [\theta_1 \ldots \theta_\zeta]; \quad \overline{H}(\theta) = \frac{1}{n_r} \sum_k H_k(\theta). \tag{20}$$

2.4.3. The salience of a grouping

Suppose x_i is an arbitrary component within a grouping $\Re$. Here, x_i may be either a point or a sub-grouping within $\Re$. Then the visual salience S of a grouping $\Re$ is obtained from the following formula:

$$S(\Re) = \gamma_{CI} \sum_i S_{CI}(x_i) + \gamma_O \sum_i S_O(x_i) \tag{21}$$

where γ_{CI}, γ_O are the weighting coefficients for the colour-intensity, and orientation salience contributing to the grouping salience. $\sum_i S_O(x_i)$ is computed from the primary oriented components of grouping $\Re$ but not from the shape of $\Re$ itself. The shape distribution or boundary of a grouping may be arbitrary and may conflict with orientations of the components in the grouping. This causes some uncertainty about how to evaluate the direction of a grouping. Here we employ a simple statistical method to deal with this problem (See [13] for other complex statistical methods involved in this field). Suppose that $x_{i_0}, \ldots, x_{i_j}, \ldots, x_{i_{n_0}} \in \Re$ are components of a given grouping with orientation components $\theta_0, \ldots, \theta_j, \ldots, \theta_{n_0}$ respectively. $C_O(x_{i_j}; \theta_j)$ is the orientation salience of x_{i_j} with orientation θ_j, $\widehat{O}$ denotes the primary orientation on which (orientation) map the grouping $\Re$ has the maximum sum value at the current layer of the orientation pyramids. A simple method to compute $\widehat{O}$ is: calculate the value sum on each θ_j orientation map of all components within $\Re$ to obtain a distribution histogram of different oriented vectors (as the horizontal ordinates); then take the orientation which has the maximum value in the histogram. The formula for calculating $\sum_i S_O(x_i)$ is then:

$$\sum_i S_O(x_i) = \sum_i C_O(x_i) \quad \text{when } \theta_j = \widehat{O}. \tag{22}$$

The above formulae for the salience computation of a grouping is a practical implementation based upon the theory discussed in Eq. (7). As mentioned before, some other factors influencing salience are not considered at the moment, for example, the relative size factor between a grouping and its surrounding groupings. When the size of a target is different from the surrounding distractors but shares all other properties with these distractors, the target will “pop-out”. The current computation method is inapplicable in this special case. This factor looks like very simple and seems easy to implement but it is not in practice. There are a lot of problems associated with it and some are difficult to resolve. One problem is how to evaluate the homogeneity of the target’s surround, especially to surrounding objects or regions. The homogeneity of a surround is affected by many factors such as shape, orientation, or colour. The shape of an object or a region may be arbitrary, so the “pop-out” by the relative size factor would depend upon the shape factor as well even if excluding other factors such as how to quantify the relationship between salience and the relative size.

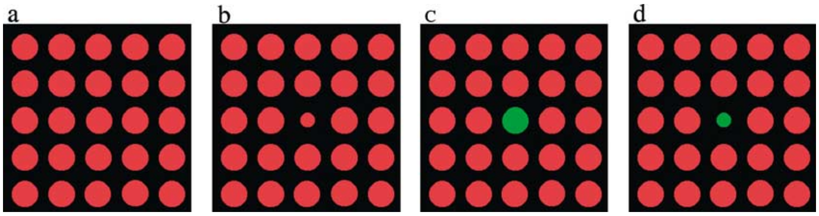

Fig. 4. An example for salience varying with relative size between the center target and surrounding distractors.

Another problem is how to evaluate the degree of homogeneity and heterogeneity of the surround of a grouping, especially under the consideration of orientation. The method (formulae (18), (19), and (20)) used in this paper is simple and may work under many homogeneous or heterogeneous environments. For example, the homogeneity surround: all neighbours with the same orientation should be different from another homogeneity surround: some different neighbour rings have (some) different orientations but on each ring the neighbours have the same orientation. But this method is not complete especially when the surround consists of arbitrary objects. As mentioned above, an object has a shape and the shape may be arbitrary. Even if ignoring other factors such as colours, how to calculate an object orientation is not easy and this directly affects the homogeneity of a surround. The difficulty is that there is no reference which can be used to evaluate an exact order of the different homogeneity distributions of orientations. Solutions for the above problems need more evidence from other research fields such as psychophysics and neuroscience.

Fig. 4 shows an example about the relative size factor. In Figs. 4(a) and (b), the red target "pops-out" in (b) when it becomes smaller. But in Figs. 4(c) and (d), which green target is more salient? Although (c) and (d) are the same to (a) and (b) except the target's colour, it may be that the target in (c) is more salient than the target in (d).

2.5. Competition pool of attention

In this module, different groupings are dynamically formed on different layers of pyramids and compete for attention selection from the coarsest level to the finest level by visual saliency interacting with top-down attentional biasing. The output is the dominant signal of the competitive winner(s) which is used to control the preferential processing or selectivities of visual attention. According to [14,15,23], the competition for visual attention can occur at multiple processing levels from low-level feature detection and representation to high-level object recognition in multiple neural systems. Also, "attention is an emergent property of many neural mechanisms working to resolve competition for visual processing and control of behaviour“ [14]. The above studies provide the direct support for the integrated competition for visual attention by binding object-selection, feature-selection and space-selection. The grouping-based saliency computation and hierarchical selectivity process proposed here, therefore, is a possible approach for achieving this purpose. Hierarchical selectivity operates on bottom-up visual salience from various groupings on each pyramid layer in the space-time context and top-down attentional setting. The outline of top-down attentional setting logic is shown in Fig. 5. It

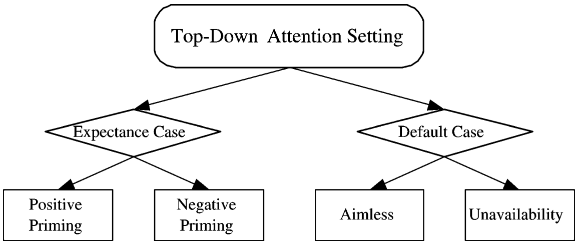

Fig. 5. Top-down attention setting.

is implemented as a control set of four attentional states for the current bottom-up visual input at any competitive moment:

(1) *Positive priming* by which consistent bottom-up input will gain a competitive advantage;
(2) *Negative priming* which is contrary to positive priming;
(3) *Aimless* or *free* state in which visual attention presents a neutral state to any visual input and thus the competition for attention is completely decided by bottom-up visual saliency;
(4) *Unavailability* state in which visual attention is occupied at the moment. It means no visual attention is available.

As pointed out in [33,50], top-down priming and bottom-up visual saliency both play important biasing roles in attention capture. Top-down biasing signals affect the competition for selective attention by increasing or decreasing the baseline of neural activity. Until sufficient psychophysical findings are found to show how top-down influence directly amplifies or reduces the intrinsic salience of targets, it is feasible to take the top-down setting into the threshold of attentional competition as proposed below. If employing a competitive neural network such as a WTA (winner-take-all) network, a top-down setting could be implemented by installing a dynamic threshold for neuron firing but the overall computational cost for dynamic attention competition is expensive. A complex structure with an enormous number of neurons with population competition is needed.

The solution presented here is to implement attentional setting via a threshold at the decision-points in the hierarchical selectivity process. Top-down attention setting here plays two roles: one is top-down biasing for globally and locally attentional competition; another is an intention request of whether to “view details” of a grouping (e.g., its sub-groupings) when attention is deployed at a grouping. However, top-down priming for special objects or groupings is very complicated since intricate object recognition from higher level processing is at least required. At present, top-down biasing here aims to act only on the level of basic features which here are the colour, intensity, and orientation feature pyramids.

The top-down signals (Table 1) include two flags for colour (which also includes intensity) and orientation top-down biasing and one flag for “view details”. Each flag encodes the states of its correspondingly top-down signal. For colour and orientation flags, “00” is the default case in which all groupings compete for visual attention in the pure

Table 1
Top-down attention setting to the basic features

colour flag	colour input	orientation flag	orientation input	“view details” flag

bottom-up way; “01” encodes positive priming in which all groupings with the positively primed feature preferentially compete for attention and at the same time other competitors are suppressed; “10” encodes negative priming which is the inverse to positive priming; “11” is the unavailable state in which all groupings having these features are prevented from attracting visual attention. For the “view details” flag, “0” signals “continue” to explore details of a grouping (i.e., its sub-groupings if they exist at the current resolution or the finer resolution) and “1” means “shift” attention from the current winner to the next potential winning grouping. The next winner will be generated from the unattended groupings at the same resolution as the current winner if these groupings exist, otherwise from the unattended groupings that lie at the same coarser resolution as the parent grouping of the current winner (see hierarchical selectivity below). This process links from the “lineal chain” to the “collateral chain”.

Hierarchical selectivity operates on the interaction between grouping salience and the top-down attentional setting at any competitive moment. The competition for visual attention occurs first among the coarsest groupings (existing at the coarsest resolution) by global competition. Through a WTA (Winner Take All) mechanism, visual attention is firstly deployed to the winning competitor. Then, a top-down or goal-driven (request) control of whether “continuing” to view the details within the current grouping or “shifting” attention out of this grouping takes place.

If switching attention, the next winning competitor gains visual attention with the aid of an “inhibition of return” mechanism which prohibits attention from instantly returning to a previously attended winner. The priority order for generating the next potential winner is:

(1) The most salient unattended grouping that is a sibling of the current attended grouping. This winning grouping has the same parent as the current attended grouping and both lie at the same resolution.
(2) The most salient unattended grouping that is a sibling of the parent of the current attended grouping, if the above winner cannot be obtained.
(3) The backtracking continues if the above is not satisfied.

Temporary inhibitions to the attended groupings can be used to implement inhibition of return. More elaborate implementations may introduce dynamic time control into different winners so that some previously-attended groupings can be visited again. But here we are only concerned that each winner is attended once.

If continuing to check the current attended grouping, the competition for attention based on bidirectionally bottom-up and top-down interaction by local competition is triggered firstly among the sub-groupings that exist at the current resolution and then among the sub-groupings that exist at the finer resolution. This indicates that the sub-groupings at the finer resolution do not gain attention until their siblings at the coarser resolution are attended. By the force of WTA, the most salient sub-grouping wins visual attention.

After attention has been directed to the winning grouping/sub-grouping, the same (top-down) goal-driven method is used to determine whether or not to "continue" to look into the details within this grouping/sub-grouping. If not, another attention "shift" takes place. If continuing to examine the particulars of this grouping/sub-grouping, another local competition triggers. When "continuing" to check an attended grouping/sub-grouping is requested, if there is no sub-grouping existing at the current or a finer resolution, hierarchical selectivity goes back to the parent of the current attended grouping. At this moment, the same "continuing/shift" attention occurs. This "continuing/shift" recursive procedure continues until the desired goal is reached or all groupings in a scene are attended.

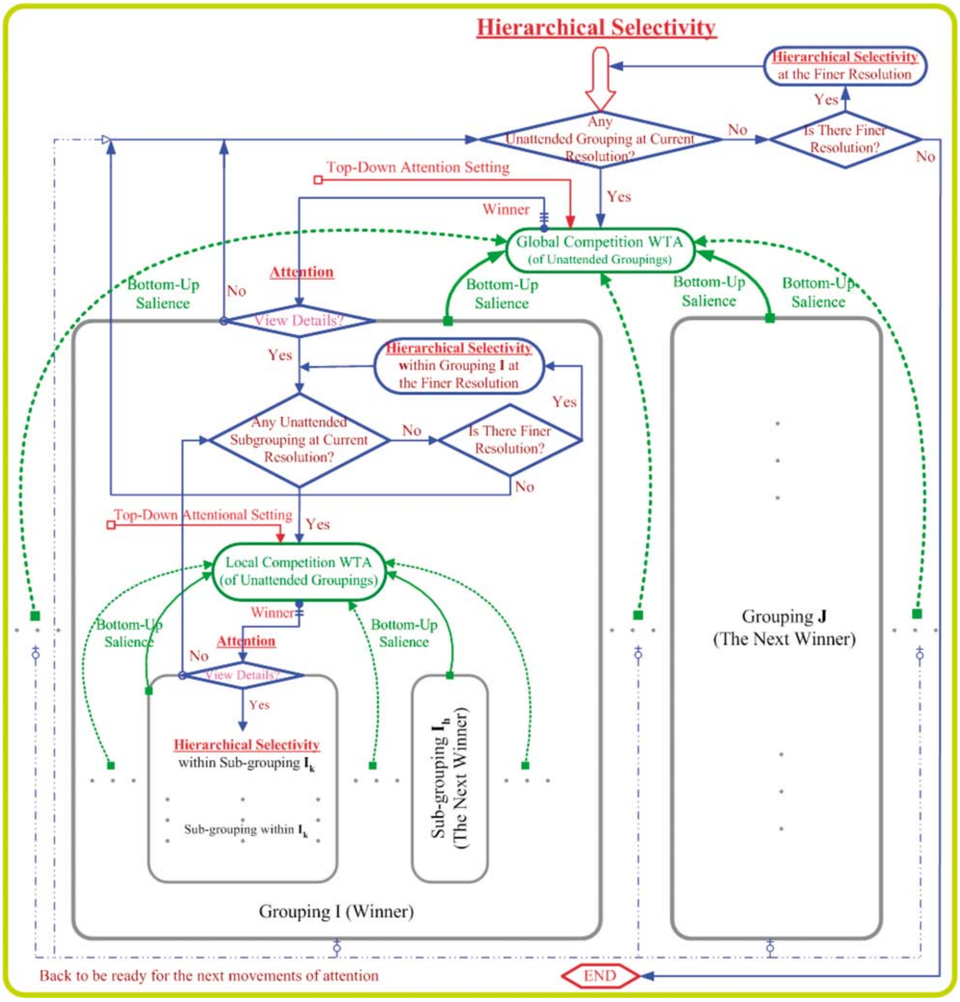
Hierarchical Selectivity
Hierarchical Selectivity at the Finer Resolution
Yes
Any Unattended Grouping at Current Resolution?
No
Is There Finer Resolution?
No
Top-Down Attention Setting
Winner
Yes
Global Competition WTA (of Unattended Groupings)
Attention
Bottom-Up Salience
No
Bottom-Up Salience
Bottom-Up Salience
Bottom-Up Salience
View Details?
Yes
Hierarchical Selectivity within Grouping I at the Finer Resolution
Yes
Any Unattended Subgrouping at Current Resolution?
No
Is There Finer Resolution?
No
Top-Down Attentional Setting
Yes
Local Competition WTA (of Unattended Groupings)
Winner
Bottom-Up Salience
Attention
Bottom-Up Salience
No
View Details?
Yes
Hierarchical Selectivity within Sub-grouping $\mathbf{I_k}$
Sub-grouping within $\mathbf{I_k}$
Sub-grouping $\mathbf{I_h}$ (The Next Winner)
Grouping I (Winner)
Grouping J (The Next Winner)
Back to be ready for the next movements of attention
END

Fig. 6. Diagram of hierarchical selectivity: see text for detailed explanation.

Table 2
The algorithmic description of hierarchical selectivity

```
1.  competition begins among the coarsest groupings at the coarsest resolution;
2.  if (no unattended grouping exists at the current resolution) goto step 10;
3.  unattended groupings at the current resolution are initialised to compete for attention
    based on their salience and top-down attentional setting;
4.  check the colour-flag and orientation-flag and apply corresponding top-down processing
    to modify the active states of the groupings (details are not implemented in this paper);
5.  all (modified) groupings compete for attention;
6.  attention is directed to the winner (the most salient grouping) by the WTA rule;
    set “inhibition of return” to the current attended winner;
7.  if (the desired goal is reached) goto step 12;
8.  if (“view details” flag = 1) (i.e., don’t view details and shift the current attention)
        { set “inhibition” to all sub-groupings of the current attended winner; }
    if (the current attended winner has unattended siblings at the current resolution)
        { competition starts between these siblings; goto step 2 and replace the grouping(s)
          by these siblings; }
    else goto step 11;
9.  if (“view details” flag = 0) (i.e., continue to view the details of the current attended
    winner)
        if (the current attended winner has no sub-grouping at the current resolution)
            goto step 10;
    else { competition starts between the winner’s sub-groupings at the current
           resolution; goto step 2 and replace the grouping(s) by the winner’s
           sub-groupings; }
10. if ((a finer resolution exists) and (unattended groupings/sub-groupings exist at the
    finer resolution))
    {competition starts on groupings/sub-groupings at the finer resolution; goto step 2;}
11. if (the current resolution is not the coarsest resolution)
        { go back to the parent of the current attended winner and goto step 2; }
12. stop.
```

As we mentioned before, the grouping salience computation is independent of how to segment the groupings in a scene (Section 2.4). The mechanism for hierarchical selectivity is also independent of what/how a segmentation is used at multiple resolutions or a single resolution for a scene. The choice of segmentation or grouping method is not included in these two mechanisms. Hierarchical selectivity runs on a given segmented scene and is driven by both the top-down attentional setting and the current distribution of the given segmentation and the corresponding salience. Switching attention between groupings/sub-groupings (and between the coarse and fine resolutions if multiple resolutions are used) is then controlled. A diagram summarizing the recursive procedure for hierarchical selectivity is given in Fig. 6. Its algorithmic description is given in Table 2.

Two goals can be achieved by taking advantage of hierarchical selectivity. One is that attention shifting from a grouping to another and from groupings/sub-groupings to sub-groupings/groupings can be easily carried out. Another is that the model may simulate the behaviour of humans observing something from far to near and from coarse to fine. Meanwhile, it also easily operates at a single resolution level. In addition, a declaration we made here is that the top-down attentional setting in hierarchical selectivity is not completely implemented in this paper although its possible approach is given in the algorithm. Except of “colour-flag = 00”, “orientation-flag = 00” and “view details flag”, other cases will be realized in the future.

Support for this approach to hierarchical selectivity has been found in recent psychophysical research on object-based visual attention. It has been shown that features or parts of a single object or grouping can gain an object-based attention advantage in comparison with those that are separated from different objects or groupings. Also, visual attention can occur at different levels of a structured hierarchy of objects at multiple spatial scales. At each level all elements or features coded as properties of the same part or the whole of an object are facilitated in tandem (see [4] for a review of these viewpoints and detailed findings).

2.6. Perceptual grouping

It has been suggested [4] that grouping processes and perceptual organization play an integral role in object-based attention. Features that are grouped together compete against other feature groupings and obtain faster processing than features that do not belong together. Perceptual grouping is a complex combinatorial problem which involves in a lot of influence factors including top-down interference in many conditions. These factors work together to affect how groupings are segmented, such as spatial proximity, similarity, common fate, shared properties, and even experience and learning [67, pp. 257–309]. In many cases, the rules for segmentation and interpretations of groupings are associated with visual tasks and experience. Nevertheless, study of this field is out of the current scope of our research. The groupings used in this paper are produced by manual preprocessing based on Gestalt principles and heuristic knowledge, to provide the basis for experiments with our attentional model. The principles of grouping used are some common rules: proximity, closure, continuity, common fate, familiarity, and shared properties. A visual grouping is defined as an effective hierarchical structure formed by all components according to these principles. For example, objects which share a common colour or orientation and are separated from their surrounding which does not share this colour or orientation may be organized as a grouping. Objects belonging to a large group or share the same spatial location may be segmented into a multi-level structured grouping. In Fig. 14, the white stripes in the road are grouped into three groupings by their familiarity. The four cars are organized as a grouping by their common fate. Two people are grouped together by their proximity.

In fact, the “grouping” we are addressing here is the “perceptual units” which serve as the potential units of attention. For object-based attention, it is the “proto-objects” produced by various segmentation processes rather than the conceptual or recognizable “objects” we commonly experience in real word. “Evidence suggests that ‘object-based’ attention and ‘group-based’ attention may reflect the operation of the same underlying attentional circuits” [76]. One general criticism of object-based attention is the question whether objects are recognized before or after the selectivity by attention or visual segmentation processes occurs with attention/without attention, that is, also the traditional story in terms of “early selection” and “later selection” or the degree of preattentive processing in the visual systems. The issues we stress here may lead to clear this misunderstanding and the detailed discussion of issues can be found in [17,76].

3. Results and discussion

For the evaluation of our object-based attention model, we ran a number of experiments based on synthetic and natural images.

3.1. Performance in synthetic images

The goal of the experiments in this section is to compare the performance of our model with human behaviour in visual attention experiments. The experiments are designed for this purpose. Additional experiments can be found in [81].

3.1.1. Neighbourhood influence on a grouping

Many psychophysical studies of visual attention (especially on object-based attention) have suggested that visual search is greatly affected by the attribute distribution and interaction between target and its surroundings (see [4,69,92] for a detailed explanation). These effects are clearly observed in experiments testing similarity or shared feature dimensions between target and non-target and homogeneity or heterogeneity of non-targets themselves. When the distractors surrounding the target are more homogeneous to each other and share less features with the target, search becomes more efficient or accelerated. Perceptual grouping also plays an important role, by which distractors are grouped by type so stronger grouping strength leads to easier pop-out [19,54,58].

We designed three kinds of experiments to test the model performance. The experiments probe the salience variation of the target in response to the surrounding changing without top-down attentional priming. It is also not necessary to calculate the target's salience on all resolution levels. For a demonstration, it is sufficient to compute the target's salience on the coarsest resolution and set the top-down attentional setting to the free-state by default. (This kind of consideration runs through the following synthetic experiments by default.)

Experiment 1 (*The scaling effect of uniform neighbours*). The experimental method is that the target is located at one place and kept fixed. Then we add more and more homogeneous neighbours which have at least one feature different to the target. The goal of this test is to prove that when the number of such homogeneous neighbours increases (i.e., the facilitated strength of the neighbours is stronger), the target's salience increases so that the target's pop-out becomes easier. We produced two series of sub-experiments to examine the model performance. In experiment A and B, the created images are all of 256 × 256 and the target is always a red bar located at the center of the displays. Green horizontal bars are gradually added in the neighbourhood of the target and kept homogeneous. Compared with experiment A, the target in experiment B is vertical. So, the target is different from its neighbours by only one feature of colour in experiment A, two features of colour and orientation in experiment B. Features considered in the computation of the target's salience are colour and orientation. Both distractors and background take part in the salience computation of the target. That is, the target's salience is derived from the contrast not only between the target and distractors but also between the target and background. Fig. 7 shows several created images and the results of the target's salience in these two experiments.

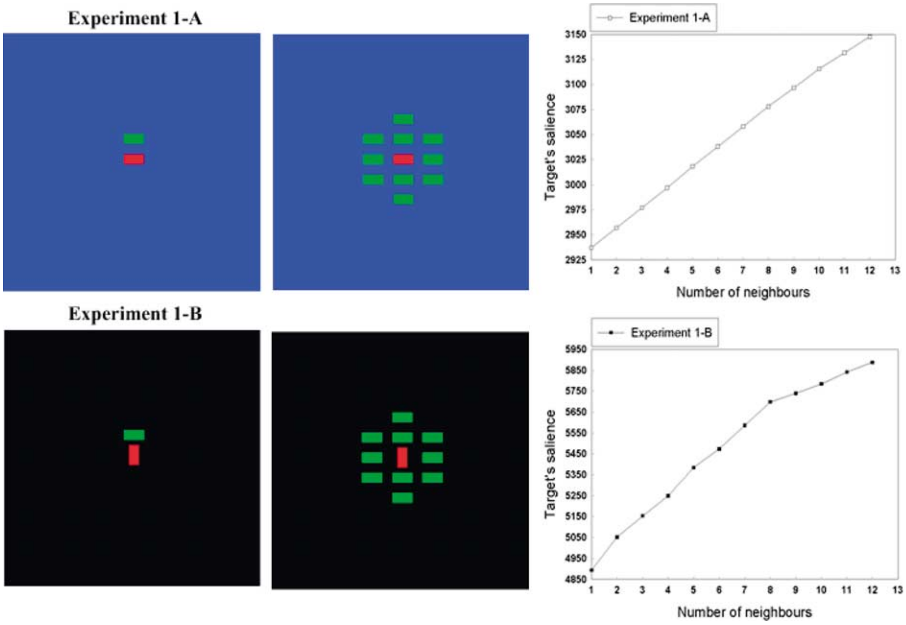

Fig. 7. The performance of the model in experiments varying the scaling effect of uniform neighbours. Left and middle columns: two of the displays used in each sub-experiment. The related results of each experiment are shown in the right graphs respectively.

Discussion. The results from experiments A and B clearly show increasing target salience with increasing homogeneous neighbours (greater strength of neighbourhood). This is consistent with the findings from psychophysical experiments. Furthermore, the curve in experiment B ascends faster than that in experiment A (notice the different scales of Y-axes in experiments A and B). It is suggested that uniform neighbours sharing fewer features with the target make the target more salient and hence attract more visual attention. We also did another experiment based on experiment A (not presented) in which we adjusted the target's size. When the target became smaller its salience became smaller. But when the target became smaller and shared the same colour with the distractors, the results became unpredictable because of the relative size factor. As we already discussed in Section 2.4.3, the model will fail to perform for this special case.

Experiment 2 (*The effect of coherence in the target's neighbourhood*). This experiment investigates the salience of the target in an originally homogeneous surrounding by gradually changing one attribute of more and more neighbours to another one (colour or orientation) but keeping them homogeneous. We produced two series of test images with size 256 × 256 for two sub-experiments. In the first sub-experiment more and more items surrounding the target change colour to be the same as that of the target. The target's salience comes from its comparison with all other circles and green background. In the second sub-experiment, the neighbour items become orthogonal to the target one by one. In this sub-experiment, the salience of the target is derived from its comparison with all

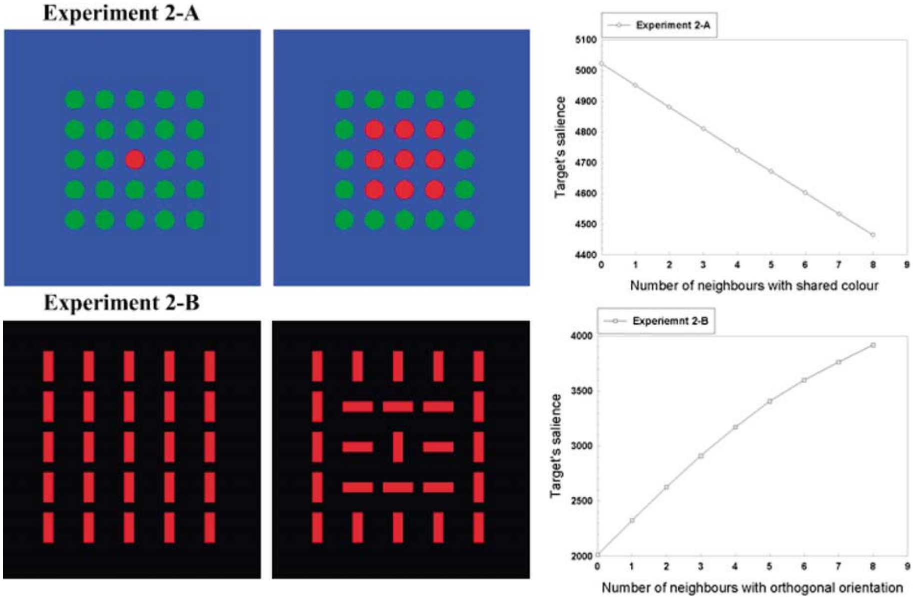

Fig. 8. Model performance when varying attributes of target's neighbours in a homogeneous environment. 2-A: the target is a red circle located at the center of the display and the neighbours change to the colour of the target. 2-B: the target is the vertical red bar located at the center of the display and its neighbours change to the orientation orthogonal to the target. Left column: first test display. Middle column: 8th test display.

other red bars and black background. To remove the effect of distance varying when a horizontal bar is rotated, the computation for distance factor is designed as: all red bars within the same neighbourhood have the same distance whatever their orientations are. That is, when a horizontal red bar is rotated to vertical, its distance remains the same as before. Several images and the results of these experiments are given in Fig. 8.

Discussion. The results shows that the target's salience becomes weaker as more neighbours share the same colour as the target in experiment 2-A, but stronger as more neighbours turn orthogonally to the target in experiment 2-B. The reason is that in experiment 2-A the strength of grouping based on the colour green within the target's homogeneous neighbourhood became weaker while the strength of grouping based on colour red is stronger. In experiment 2-B, although the neighbours form two types of groupings, the new continuously growing grouping did not affect the neighbourhood homogeneity but enhanced the contrast to the target. In fact, both experiments have the same nature but reflect different aspects of the effect of the target's neighbourhood. The result of experiments 1 and 2, as pointed out in [20,42] and other researches on object-based visual attention, have shown that stronger grouping distractors and greater differences between the target and distractors enable the target to be sought more efficiently. In other words, stronger contrast between the target and its neighbourhood makes the target more salient to capture visual attention in the bottom-up competition.

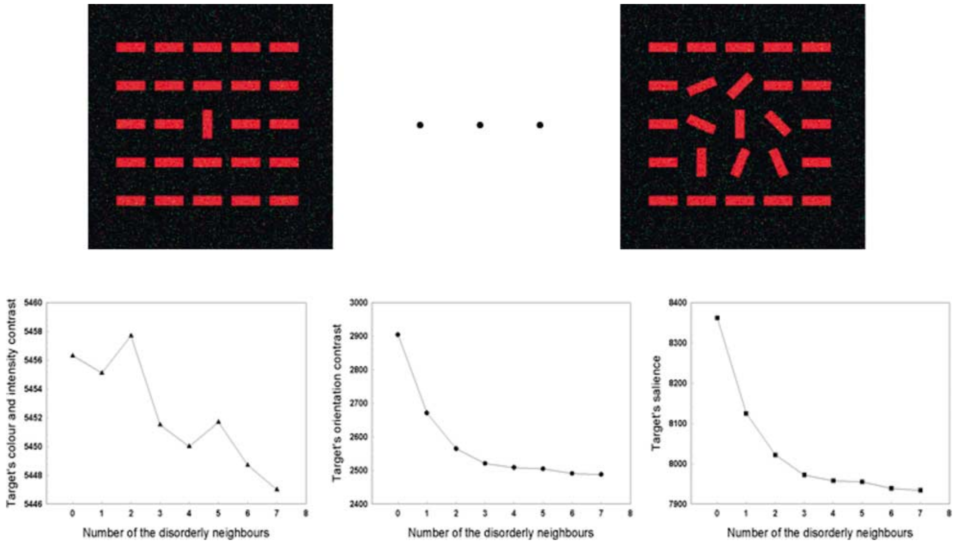

Fig. 9. Model performance in an oriented heterogeneous environment. The target is the red vertical bar located at the center of the display. Its neighbours become more and more heterogeneous by gradually varying their orientations from the target and each other. Two members of the sequential displays are shown here.

Experiment 3 (*Effect of the target neighbourhood heterogeneity*). This experiment examines the performance of the model in heterogeneous circumstances. In theory, the target should be less salient with a more disorderly distribution of the neighbourhood. The method used here is similar to the two previous experiments. The red vertical target was initially located at the center of a homogeneous surrounding in which the same colour bars are orthogonal to the target. After that, we gradually varied the neighbours' orientations to create a series of more and more heterogeneous displays. One experiment is shown in Fig. 9. All displays have added 30% random colour noise. The target's salience is computed by colour and orientation of the target contrasting with both of the distractors and background. Although we do not give results from all the experiments, the overall experimental results are similar to that in Fig. 9.

Discussion. The results shown in the bottom diagram in Fig. 9 indicate that the target's salience decreases with the growing heterogeneity of its surroundings. This means that the efficiency of visual search becomes worse and worse. Notice that the downtrend of salience is much steeper in the first four steps and tends to a mild decline afterwards. The saturated tendency effect is not surprising but expected. The Gabor filter for orientation extraction used here is sensitive to four orientations of 0°, 45°, 90°, and 135°. When the number of disorderly orientations exceeds four directions, the result is an almost saturated weight ω in equation (19) (see Section 2.4.2) because ω is limited by the maximum different orientations (4 here). This ω is used to evaluate the orientation disorder within an object's neighbourhood. Another observed phenomenon from the graphs in Fig. 9 is that the main contributor to regularly reduce the target's salience is the orientation disorder factor rather than feature colour. The explanation for this effect is that the distractors always shared the same colour with the target and the varying position of each pixel within each distractor grouping in this experiment produced only a tiny effect in the colour contrast between the target and the distractor, so the overall trend of the target's salience is hardly affected by the colour of the surrounding features.

We have also examined the behaviour of the model performance with varying target intensity with a random noise background, varying target orientation from 0 to 360 degree [65], and the eccentricity of the target location [92]. The results on these experiments are compatible with the corresponding findings from the human psychophysics literature.

3.1.2. Grouping effect and related hierarchical selection

Fig. 10 shows a display in which the target is the only vertical red bar and no one of the bars has exactly the same colour as another bar. Three bars have the same orientation and others have different orientations. If we do not use any grouping rule, each bar may form a single grouping by itself. Then we obtain 36 single groupings. If segmenting the display by shared orientation, the only structured grouping is formed by the 3 vertical bars, which includes the red target (forms one sub-grouping) and other two vertical green bars (forms another two-level sub-grouping). In this way, 34 top groupings (38 in total) can be obtained: a structured three-level grouping (contains 4 sub-groupings) and 33 single groupings formed by other distractors respectively. The resulting salience maps and attention sequences for these two segmentations are given in Fig. 10. The background, colours, and orientations are considered in the salience computation. The top-down attentional setting is set to the free state, so this is pure bottom-up attention competition.

The results show different orders of paying attention to the targets. The target belonging to a grouping (see Fig. 10 (C1), (C2), (C3), (C4), and (C5)) has an advantage in attracting attention much more quickly than the non-grouped. The competition starts among different groupings in the display. The structured grouping of 3 vertical bars is the most salient compared to others and obtains attention firstly. Then the competition occurs within this grouping between the target and another sub-grouping formed by the two vertical but different colour bars. Attention is directed to the sub-groupings according to their salience orders when we do not consider top-down attentional priming. The target is attended after the two-level sub-grouping is attended. This grouping advantage for attentional competition has been confirmed by psychophysical research on object-based attention [4, 76].

3.2. Performance in natural scenes

In the previous section we examined several aspects of our attentional model by using some artificial images and successfully compared our results with related findings in psychophysical research. To investigate the model in complex natural scenes, we used colour outdoor photographs taken with a digital camera. The implementations for both of "from coarse to fine" and "from far to near" human eye simulations in these real imagery are described in detail.

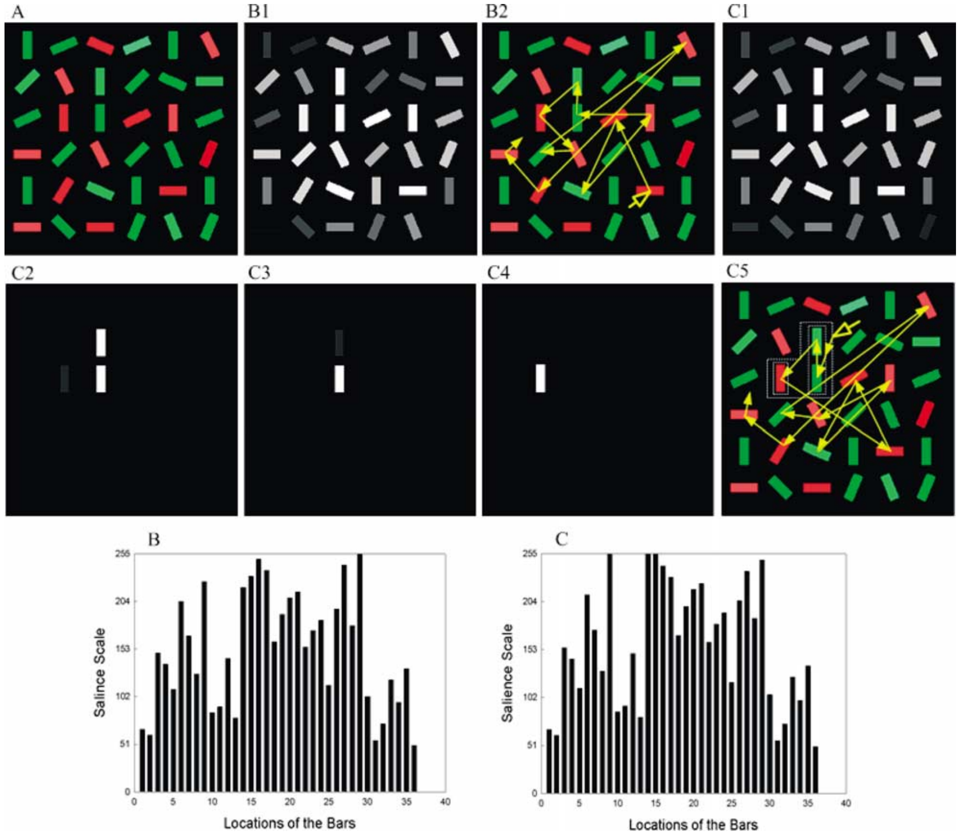

Fig. 10. An example for structured groups and hierarchical selection. In the display the target is the vertical red bar at the third row and the second column. B1: salience map (in shades of grey) in the case of no grouping. B2: partial attention sequence of most salient bars for B1. C1: salience map in the case of grouping. C2, C3, C4: salience maps of the grouped bars. C5: partial attention sequence of most salient bars for C1. B, C: salience histograms for B1 and C1 respectively. Note target is attended after 7 shifts in B2 but only 3 in C5.

3.2.1. Hierarchical selectivity

As suggested in [76], “there may be a hierarchy of units of attention, ranging from intra-object surfaces and parts to multi-object surfaces and perceptual groups”. Hierarchical selectivity is a novel mechanism designed for shifting attention from a grouping to another one or from a parent grouping to its sub-groupings as well as implementing attention focusing from far to near or from coarse to fine. It can work under both multiple (or variable) resolutions and single resolution environments. Resolutions can be either scaled by pyramid decomposition scheme or by digital camera. Here an outdoor scene is used to demonstrate the behaviour of hierarchical selectivity. In Fig. 11, the same outdoor scene is photographed from far and near distances respectively so that two coarse (64×64) and fine (512×512) resolution photographs are obtained. In the scene, there are two groupings: a simple shack in the hill and a small boat including five people and a red box within this boat on a lake. The people, red box, and the boat itself constitute seven sub-groupings

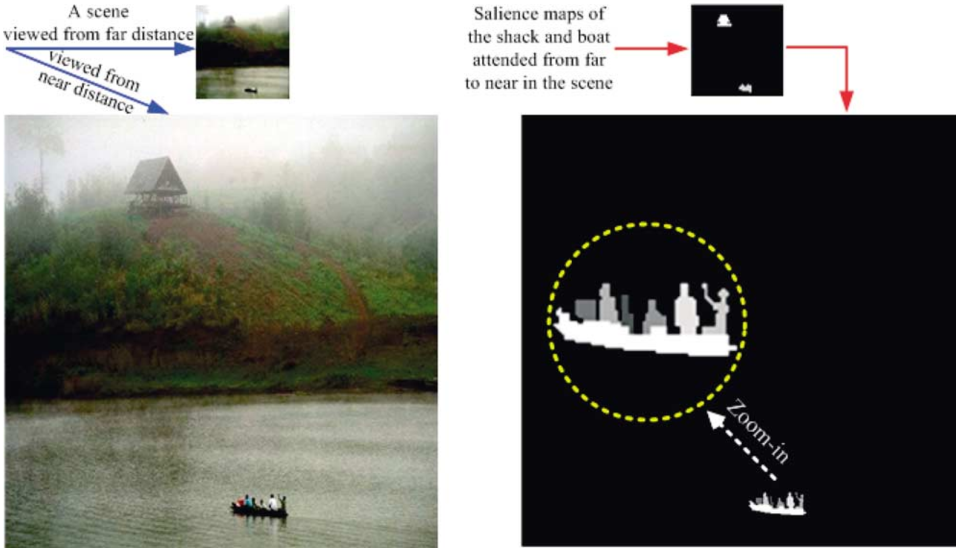

Fig. 11. An outdoor scene photographed from far and near distance respectively. The obtained images shown here are the same scene but different resolutions. The salience maps is shown too and the grey scales indicate the different saliences of the groupings.

respectively for this structured grouping. The $1/\rho$ parameter for Gaussian distance is set to 25% and the Gabor filter is sensitive to 4 orientations $[0^\circ, 45^\circ, 90^\circ, 135^\circ]$.

The model works with these two images, using the coarse and fine images as different resolution levels. For this purpose, only feature (colour, intensity, and orientation) maps at the lowest level of the pyramids are created for each image. (Multi-level pyramids used to simulate attending a complicated natural photograph from far to near and coarse to fine is also implemented in this paper. See the next sections for details.) Competition for attention starts in the coarse or far image (Fig. 12). Using hierarchical selectivity, attention is firstly deployed to the winner (here the shack) and suppresses other competitors. Then attention shifts to the fine image for further checking this winner if answering “yes” to the “view details” flag. If the answer is “no”, the model will check if there is(are) any other grouping(s) existing at this image. When attention is shifting, an “inhibition of return” is set for this attended grouping. Because the shack has no sub-groupings, attention switches again to the coarse image and checks if there exists any next winner. Thus the boat grouping obtains attention. In the same way as attending the shack, answering “yes” to the “view details” flag attention shifts to its sub-groupings in the fine image. At this moment the competition for attention triggers among the seven sub-groupings. Attention is deployed to these sub-groupings by hierarchical selectivity. The salience maps computed for these groupings are shown in Fig. 11 and the sequence of attention deployments is shown in Fig. 12. The attention deployment trajectory shown in Fig. 12 reveals reasonable movements for this natural scene.

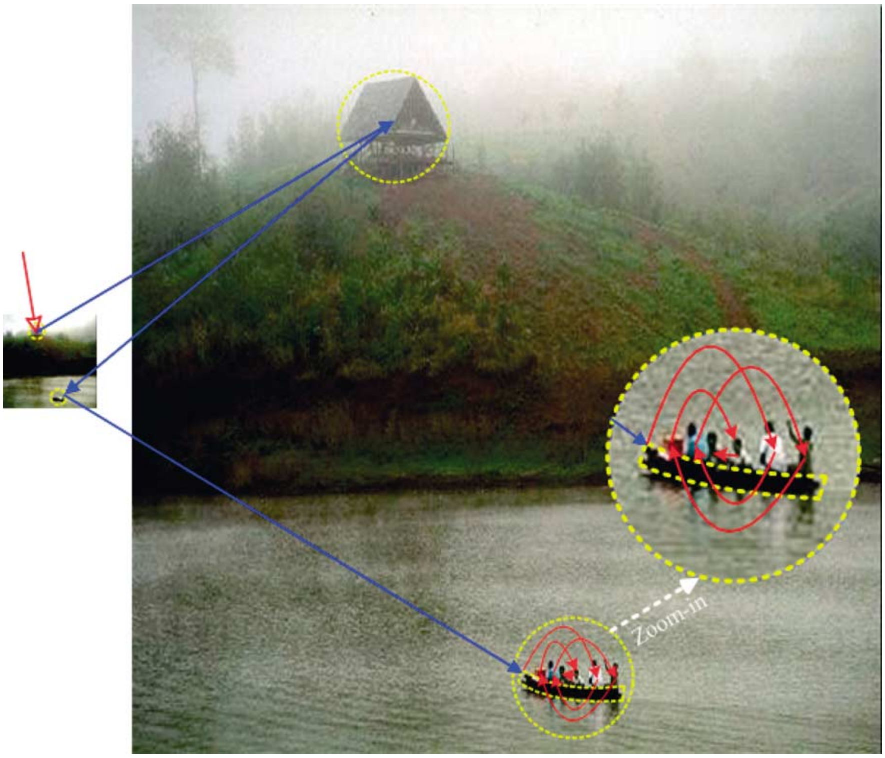

Fig. 12. The attention movements implemented for the outdoor scene: blue arrows indicate attentional movements between resolutions and red arrows denote attention shifts at fine resolution.

3.2.2. Hierarchical selectivity from coarse to fine

The image presented in Fig. 13 has 512 × 512 pixels and contains many structured objects and groupings. The pyramids in the model used here have three layers, ranging from resolutions 128 × 128 to 512 × 512. The Gabor filter was set to be sensitive to 4 orientations $[0^\circ, 45^\circ, 90^\circ, 135^\circ]$. The $1/\rho$ parameter for Gaussian distance was set to 50%. The model firstly extracted colours, intensity, and orientations from the photo and constructed altogether 9 three-layer pyramids: one intensity pyramid (Fig. 15), four colour pyramids (Fig. 16), and four orientation pyramids (Fig. 17). Eleven meaningful groupings of objects were created manually by preprocessing according to Gestalt grouping rules (see Section 2.6). Fig. 14 shows the identifiers of the different groupings in this image. The numerals pointed to by each white arrow denotes the identifier of each grouping at multiple resolutions. The groupings which have the same prefix identifier belong to the same parent grouping. The depth of each grouping is the index of its array mark. For example, identifier 1-1 indicates that this is the first sub-grouping of grouping No. 1. Identifier 1-1-2 denotes it is the second sub-grouping of grouping No. 1-1. Groupings No. 1-1-1 and No. 1-1-2 have the same parent grouping No. 1-1. The black circles or ellipses are used to conveniently distinguish different groupings (object(s) in the circles) and not the grouping boundaries. When viewing these groupings at different resolutions, some groupings/sub-groupings will

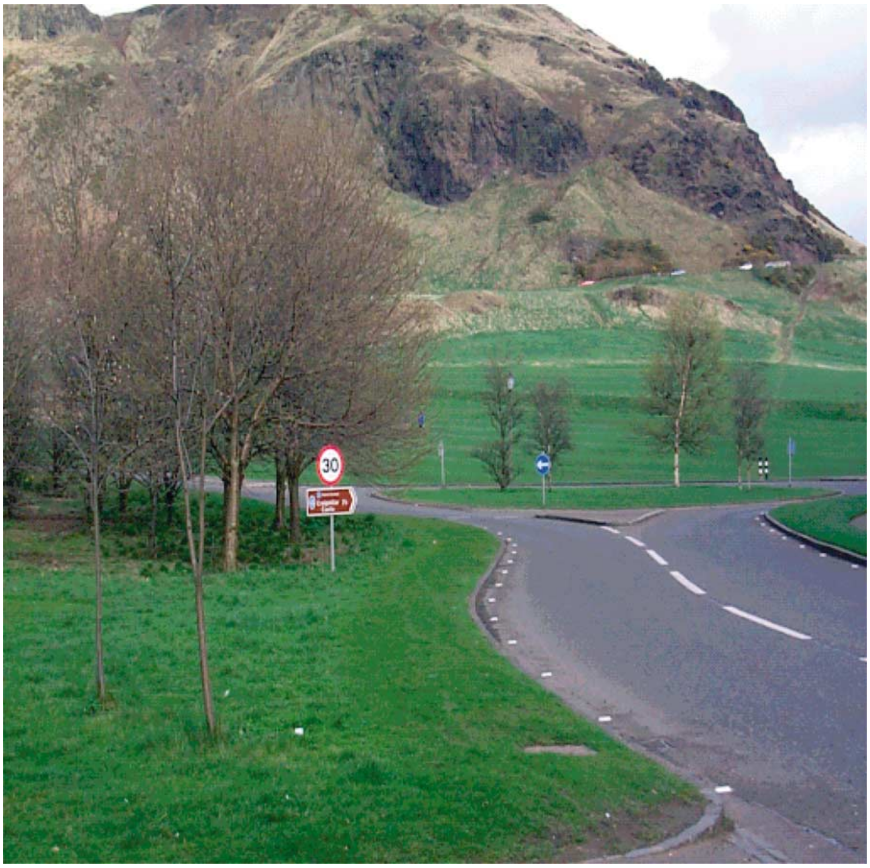

Fig. 13. An outdoor photograph.

disappear at the lower resolution. The hierarchy of groupings is shown in Fig. 14 and Fig. 26 which is discussed later.

The top-down attention setting was always set to the free state in this test. The decision-points during hierarchical selectivity to drive whether or not viewing the details within a grouping were always answered “Yes”. Although this may make hierarchical selection look like an exhaustive exploration, the general performance of the model can be inspected in detail and completely (see the next section for an alternative implementation in this view). As we discussed in Section 2.5, the control for recognizing which object is significant is very intricate and needs higher visual processing related to the current visual tasks (also see the following discussion about the small white stripes in this scene). Future work will refine this complicated control. In the more normal scenes, top-down priming proposed for the “view details” flag will control choice to produce more interesting behaviour.

Here, the competition for visual attention was firstly triggered at the coarsest resolution, namely the highest layer of the pyramids. During the attentional movements, shifting into the higher resolution (lower layers of pyramids) or switching to the lower resolution

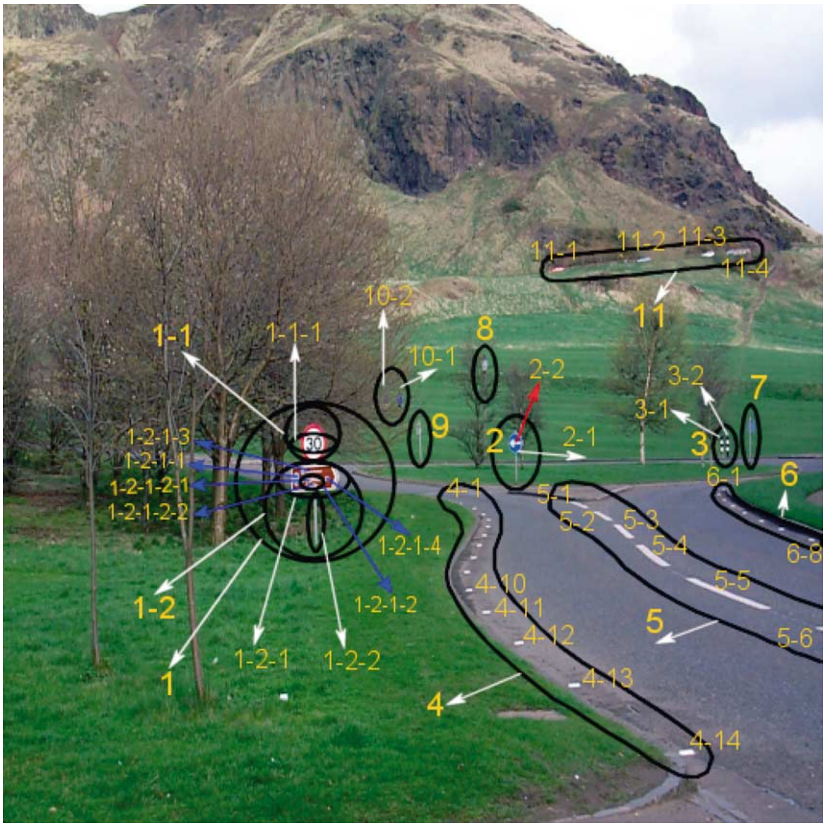

Fig. 14. The identifiers of groupings in the given photograph.

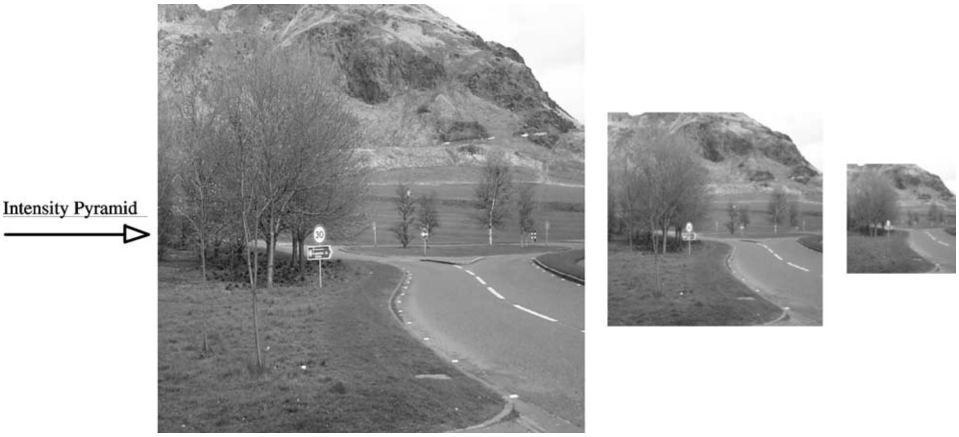

Fig. 15. The intensity pyramid built from the photograph given in Fig. 13.

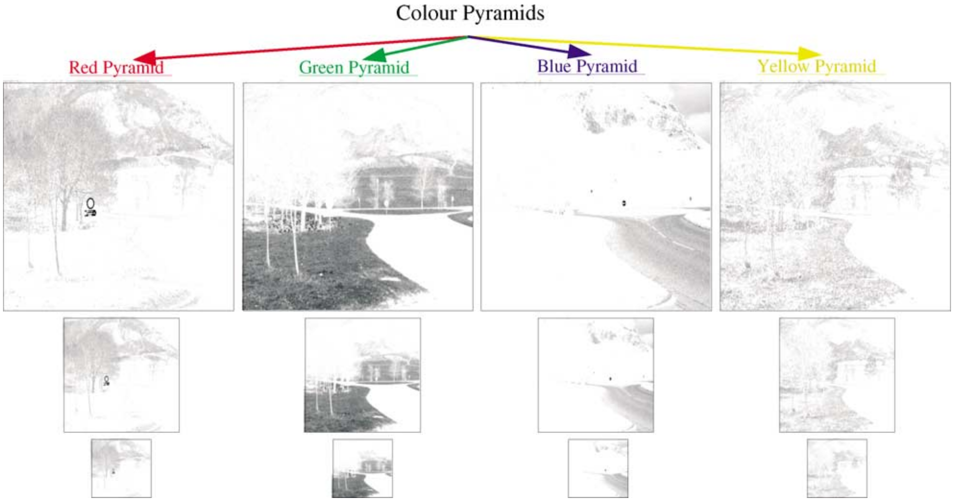

Fig. 16. The four colour pyramids built from the photograph given in Fig. 13. (The graphs are black-white inverted to improve visibility.)

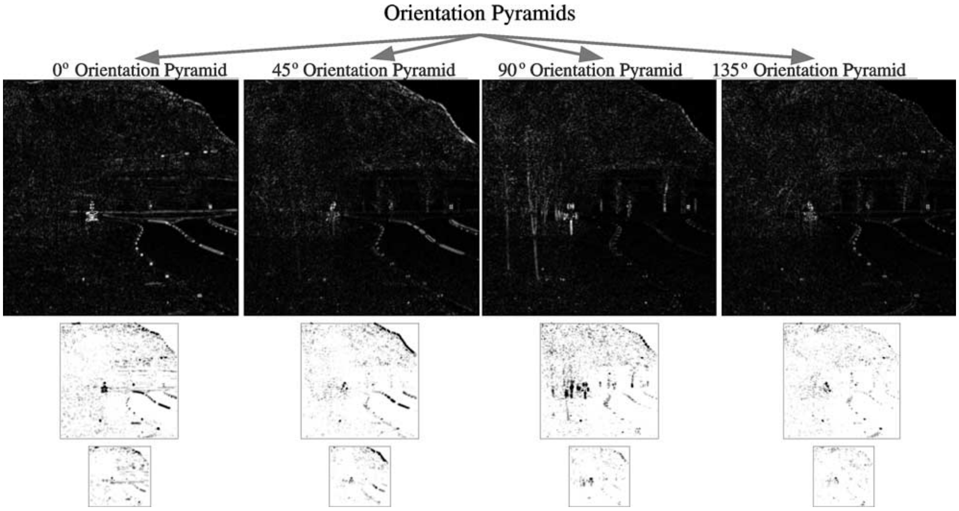

Fig. 17. The four orientation pyramids built from the photograph given in Fig. 13. (Graphs in the second and third rows are black-white inverted to improve visibility.)

(higher layers of pyramids) dynamically changed depending on the natural structure of the current grouping being attended and its surroundings. When a structured parent grouping is attended at high resolution, some/all of its sub-groupings will be attended next at this current resolution if these sub-groupings appear at the same resolution, or at the lower resolution if some/all of its sub-groupings do not appear at the current resolution. In this procedure, some sub-groupings within a parent grouping, such as some small white stripes

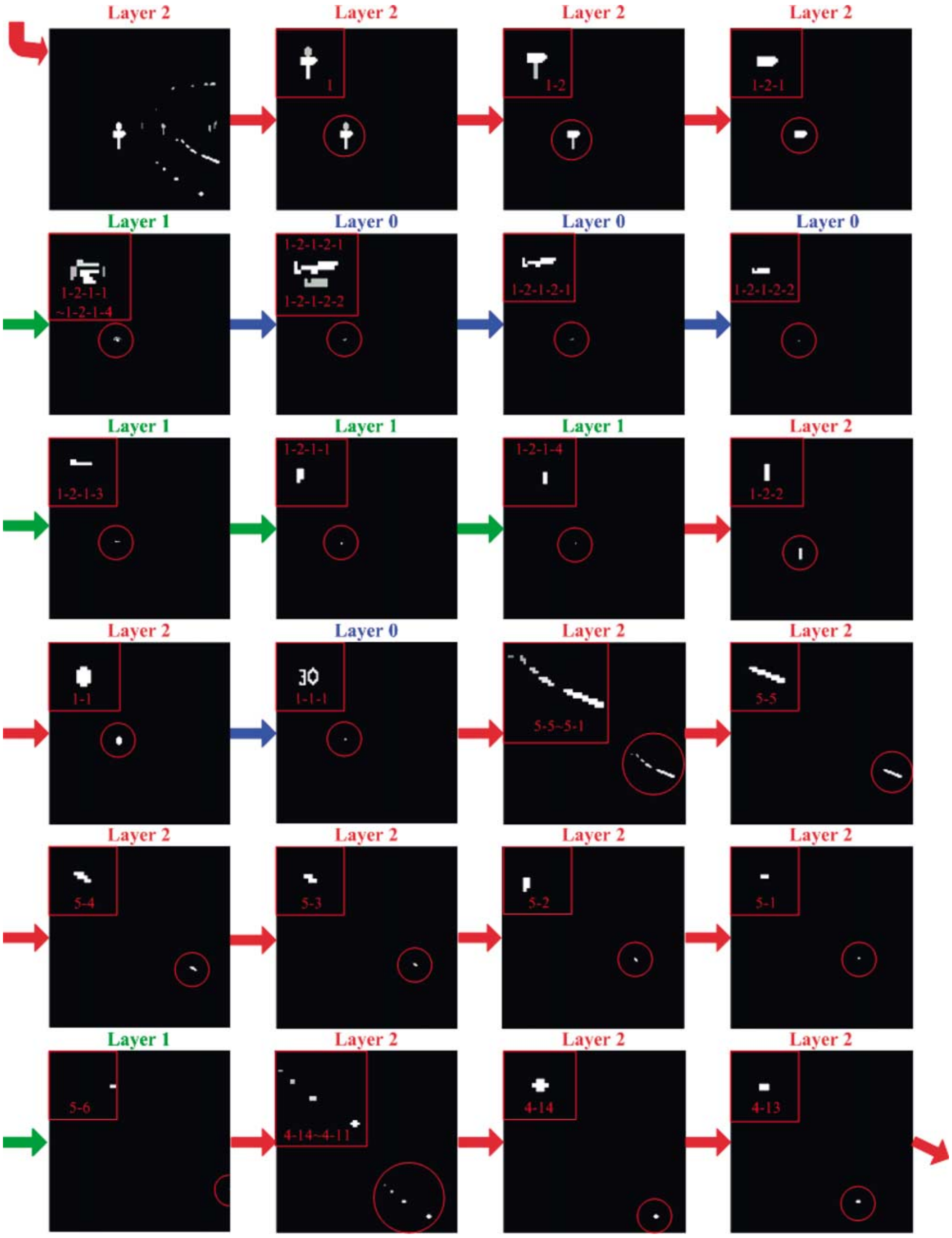

Fig. 18. Salience of the attending grouping and competing groupings, as well as the sequence of attentional movements. The red, green, and blue arrows denotes that attention is at or switched to the coarsest resolution, middle resolution, and finest resolution respectively. The small red panel at the top left corner in each slide shows a zoomed view of the objects. The red circle/semi-circle indicates the focus of attention. The grouping identifiers are also given in each panel.

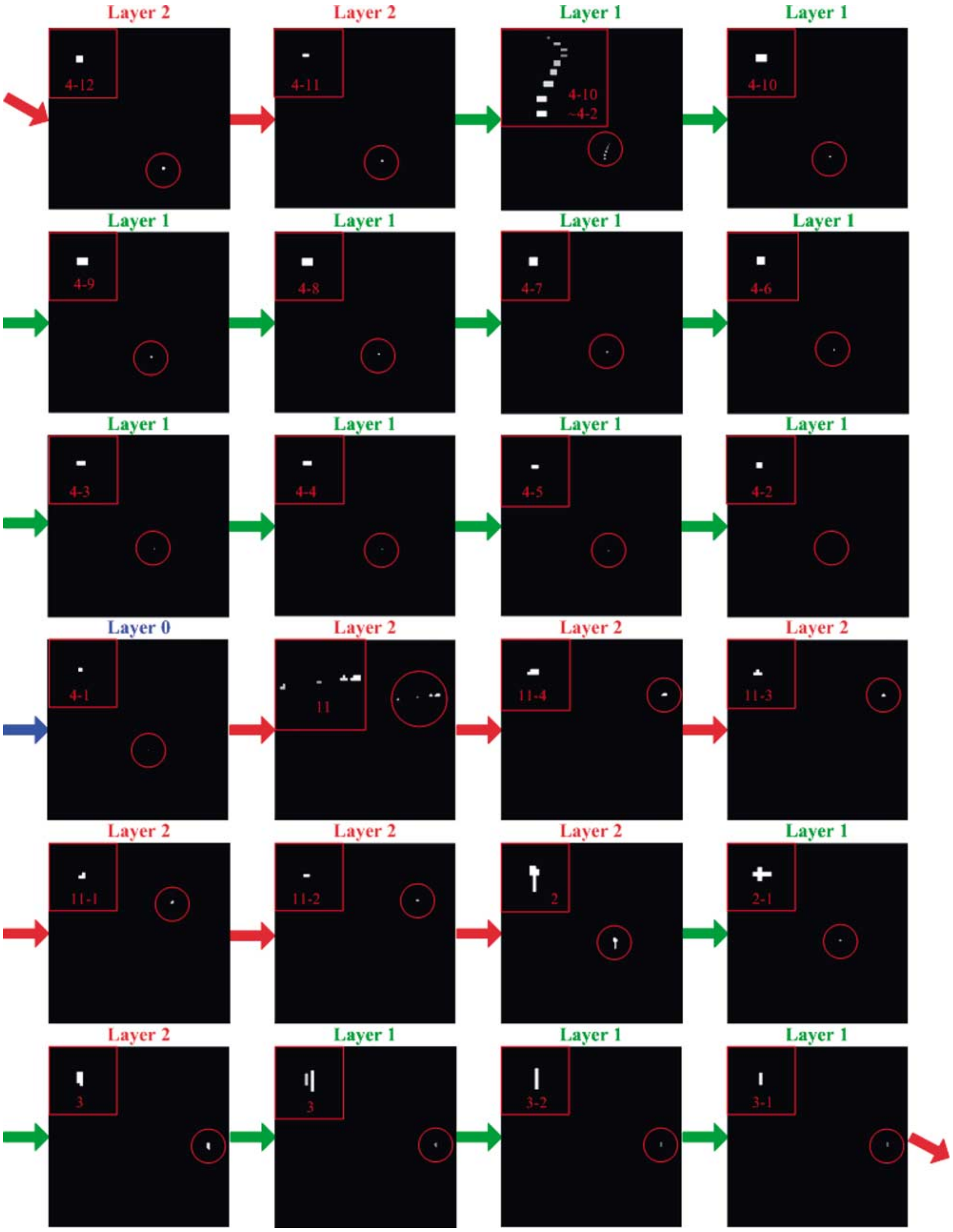

Fig. 19. Continued slides of Fig. 18.

in the road, may have not much significance and may not necessarily be attended. This further top-down control for shifting attention will need additional theory to incorporate the measure of object similarity, subject's experience and the current visual task, etc. and

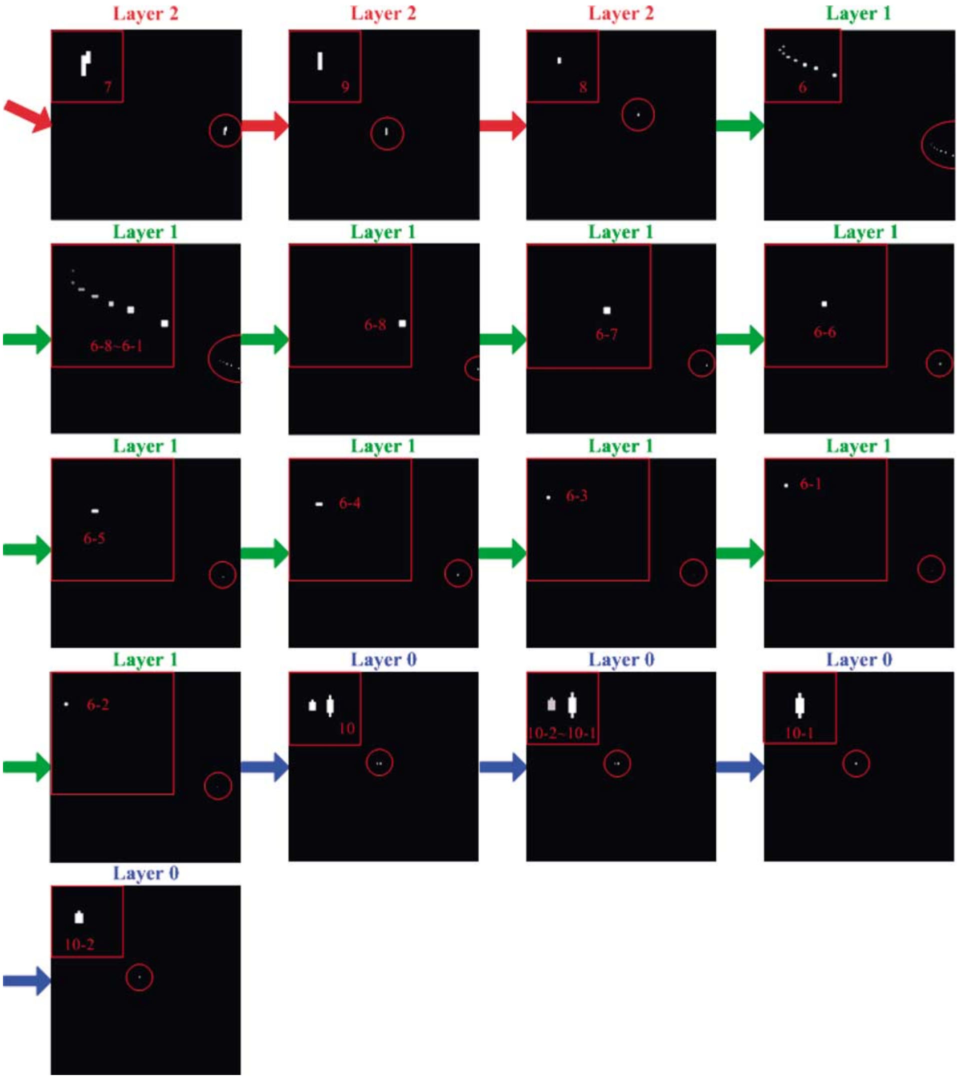

Fig. 20. Continued slides of Fig. 19.

is not implemented here. The results of the model performing on all resolution levels are shown in Figs. 18–20.

At each attentional deployment, we show the entire or unitary salience of the grouping which is currently being attended. When the related groupings are ready to compete for visual attention we present the degrees of their individual salience (in shades of grey) in comparison with all other competitors. The brighter a grouping is, the more salient it is. It is worth noting that no mosaic appearance is seen in the results because the model theory is based on object attention in which a grouping competes for attention using its

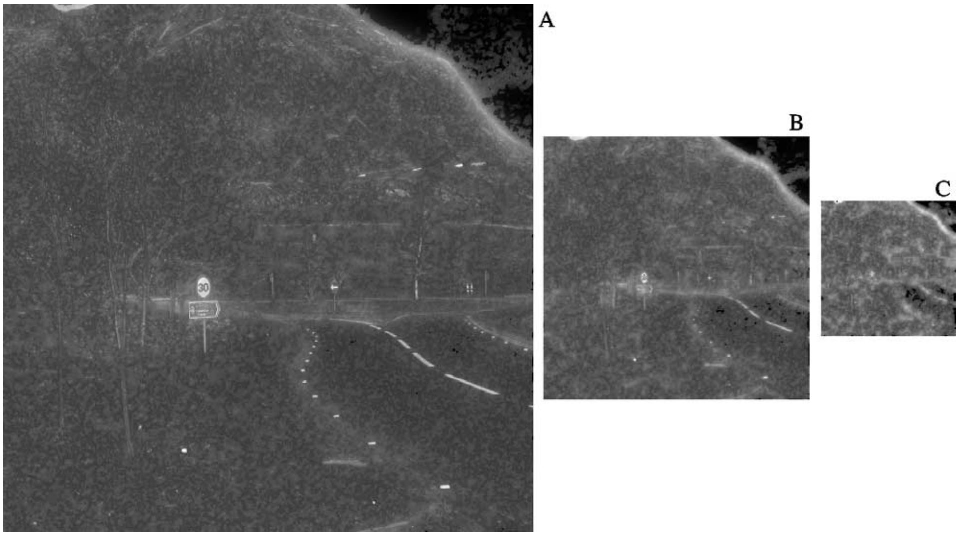

Fig. 21. Applying the model for space-based attention. Each pixel is an individual grouping. Only the raw salience maps of pixels at three resolutions are shown here in shades of grey.

entire salience integrating the strength of all its components. Thus, the salience shown is the grouping salience rather than that of each pixel within the grouping. However, our grouping-based computation approach can also be applied for spatial attention if we consider each pixel as a grouping. Fig. 21 gives the salience maps obtained from the same outdoor scene for individual pixels at the coarsest resolution (graph C), middle resolution (graph B), and finest resolution (graph A). The $1/\rho$ parameter for Gaussian distance for this experiment is set to 2%.

According to the obtained results, the order of attention shifts is shown in Fig. 22. We can see, the attention movements basically coincide with the salience difference between the objects in the scene. Some groupings, such as grouping 6, which consist of several very small sub-groupings, do not exist at the coarser resolution. They either have no way to take part in the competition, or lose much support from their smaller members or components or from their surroundings which may be useful to compete for attention at the finer resolutions. So generally, they lost some possible advantages when at the finer resolutions.

3.2.3. *Hierarchical selectivity from far to near*

Three colour images shown in Fig. 23 are taken using different resolutions from far to near distance (64 × 64, 128 × 128, and 512 × 512) for the same outdoor scene. The scene is segmented (by hand) into 6 top groupings (identified by the black colour numbers: one object grouping 6 and five regions here) and 5 of them are hierarchically structured except grouping 4. In the coarsest image, only grouping 6 (one boat including two people) can be seen. In the finer image, sub-groupings 5-1 and 5-3 within top grouping 5 appear but they lose details at this resolution. The smallest boat (i.e., sub-grouping 5-2 of grouping 5) can only be seen at the finest resolution. The salience maps of groupings during attention

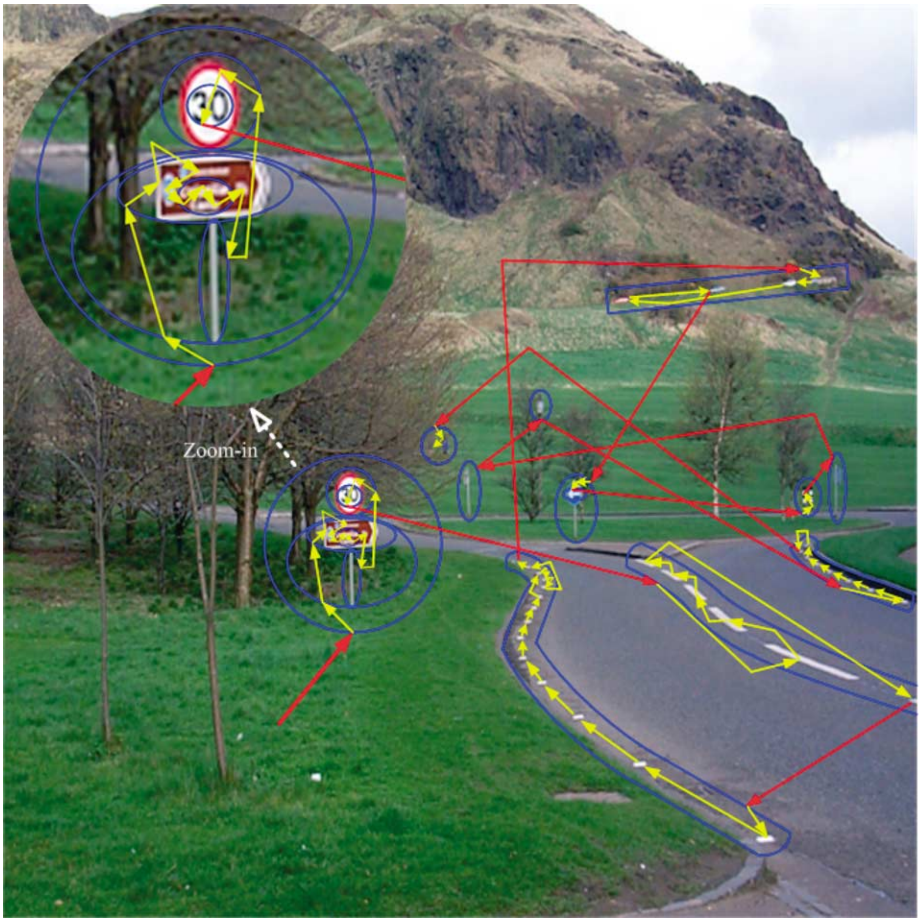

Fig. 22. The overall trajectory of attentional movements of the model in multiresolution. Red arrows show attentional shifts from one grouping to another. Yellow and purple arrows show attention switches within groupings. The circles denotes the locus of attention.

competition are also briefly shown in Fig. 23 where darker grey shades denote lower salience.

The competition first occurs among the top groupings at the coarsest scene. The most salient grouping 6 therefore gains attention. When giving a “yes” to the top-down attention setting (“view details” flag), attention will shift to the sub-groupings of 6. Two people and the boat then begin to compete for attention. If a “no” is given or after grouping 6 is attended, attention will shift to the next winner grouping 2. If a “yes” is given too to the “view details” flag of grouping 2, attention will first select sub-grouping 2-1 and then shift to sub-grouping 2-2. After attending 2-2, if continuing to view the remainder of grouping 2, attention will shift to the finer resolution to visit 2-3. When grouping 5 is attended, the lake (excluding grouping 6) is visited first and then attention shifts to the finer resolution scene where 5-1 and 5-3 start to compete for attention. In the case of giving a

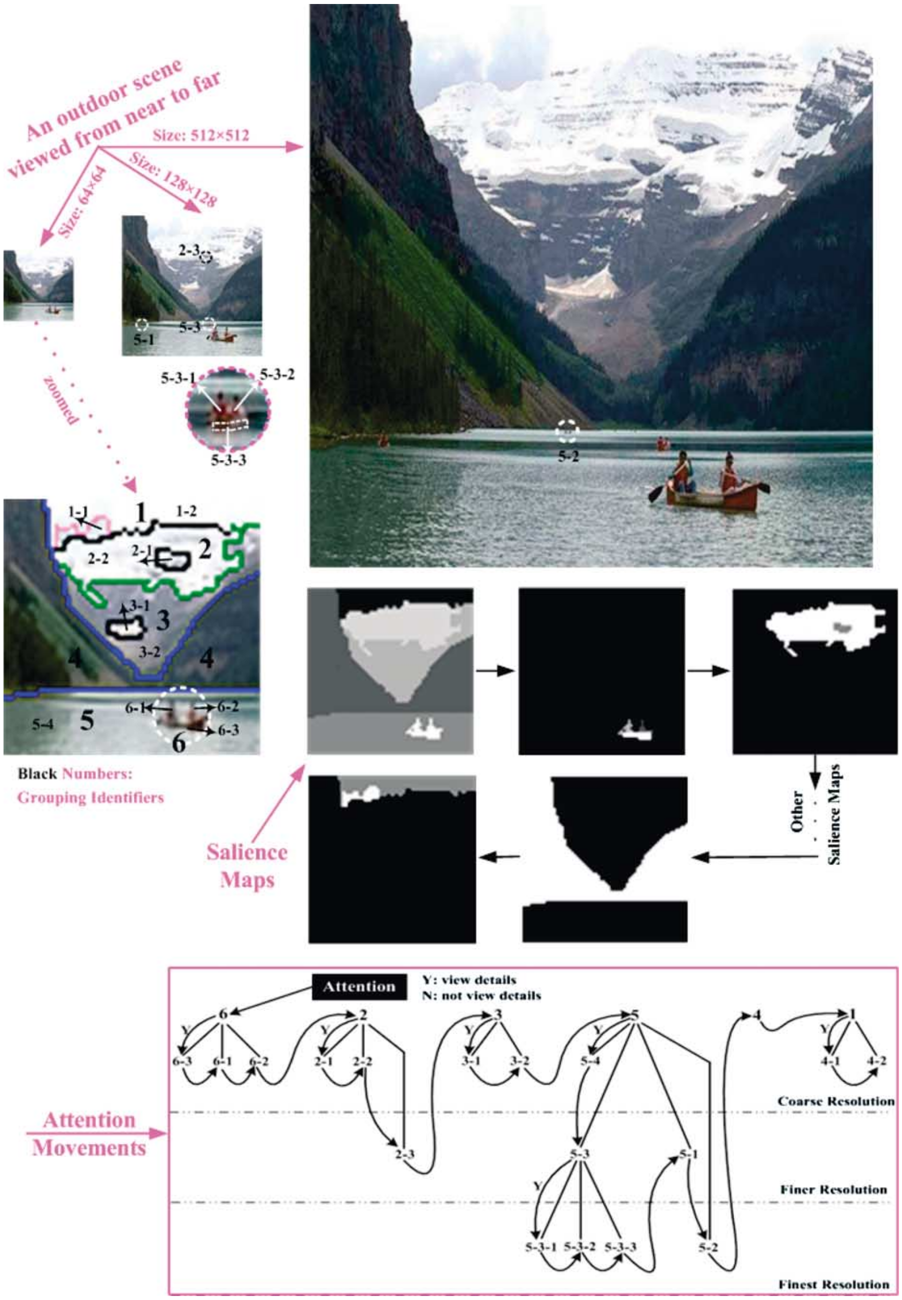

Fig. 23. An outdoor scene taken from different distances. The dotted circles are used to identify groupings but not their boundaries. The sequence of salience maps used for each selection of the next attended grouping is shown at the middle. Attention movements driven by hierarchical selectivity is shown at the bottom using a tree-like structure.

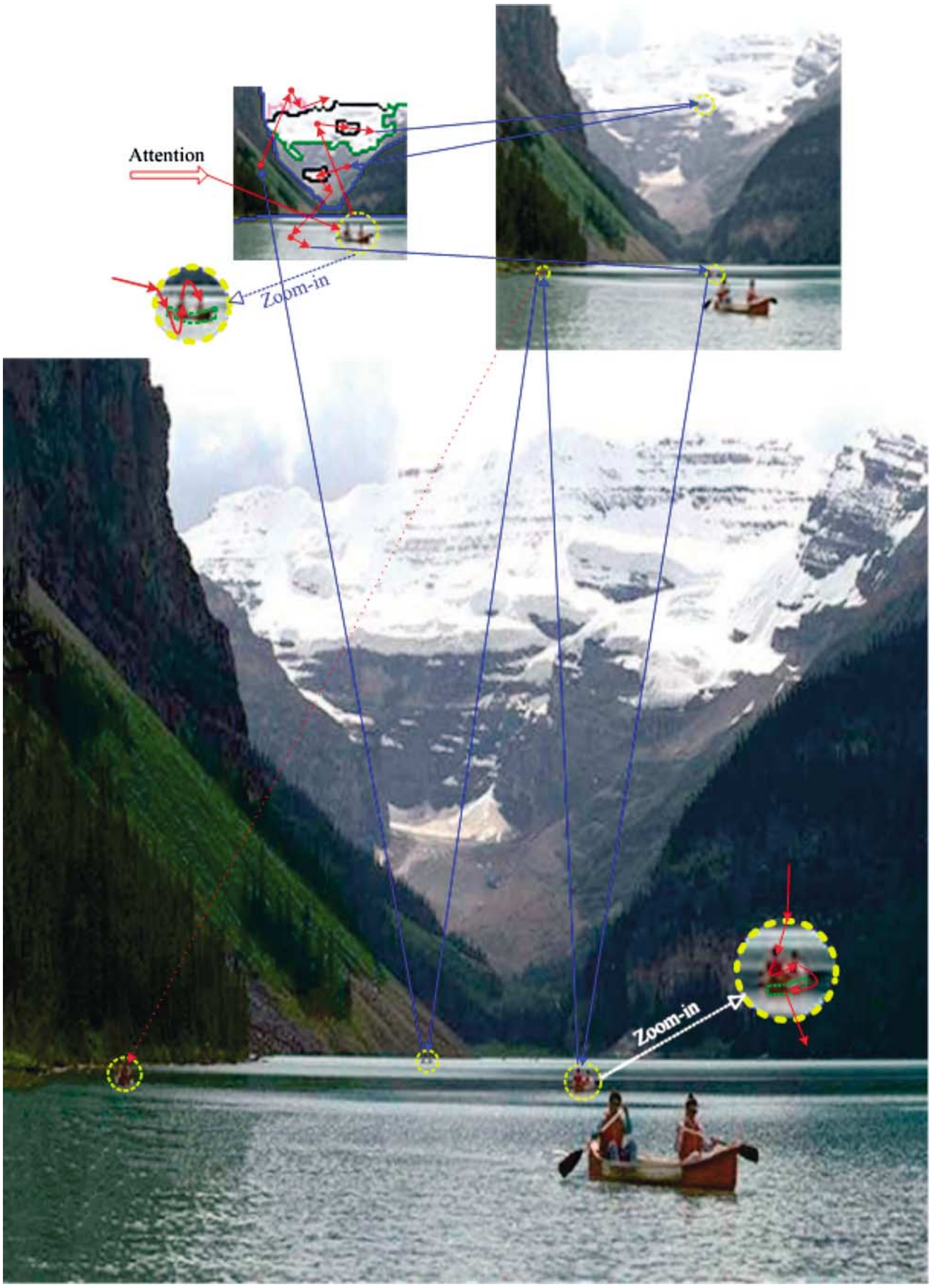

Fig. 24. The attention movements implemented for the outdoor scene: blue and red arrows indicate attention shifts between and at the same resolutions respectively. Arrows with red solid circles denote attention is attending the top groupings.

“yes” to the top-down flag of the winner 5-3, attention will shift to the finest resolution scene to check its details. Then attention goes back to the previous finer resolution scene and shifts to 5-1. After that, attention shifts again to the finest resolution scene. Thus the smallest boat 5-2 at the finest resolution is attended. Fig. 23 shows the overall behaviour of the model performed on the scene. Using this same scene, when stronger and stronger noise was added above $\sigma = 17$ for Gaussian noise, the order of the attention movements changed. The above results clearly show hierarchical attention selectivity and appropriated believable performance in a complicated natural scene. In addition, although this model is aimed at computer vision applications, the results are very similar to what we might expect for human observers. The attention movements shown in Fig. 24 reveal the reasonable shifts of visual attention for this natural scene.

3.3. Improved behaviour of hierarchical selectivity in natural scenes

We have shown the model performance in the complex natural scene. For a complete examination, we gave a positive response to each “view details” flag. However, some small stripes (on the road) may be irrelevant to the current visual task and are thus unnecessary to attend in turn. Also some tiny unreadable characters are probably not worth notice by the observer. One possible way to improve the performance on these targets is to incorporate a top-down recognition component or learning process to produce a control function with reasonable salience thresholds according to different environments and visual tasks. Our current model does not yet implement this complicated top-down control. Instead, we propose an alternative demonstration of our model’s abilities by using a simple human-computer interaction to give a positive or negative response to the “view details” top-down attentional setting (see Section 2.5 for more details).

Fig. 25 shows a logical diagram of attentional movements in hierarchical selectivity working on a hierarchical scene containing three structured groupings. In this diagram, groupings A, B, and C have a decreasing salience order and the left sub-groupings have greater salience than their right siblings. That is, the saliences of A1, A111, B1, and C1 are greater than that of A2, A112, B2, and C2 respectively. Suppose that attention is currently deployed at grouping A111 and a negative answer is given to the check flag of the top-down attentional setting “view details”. Then there are multiple (here four) possible destinations of the next attention movement, shifting to A112, A12, A2, or B (as shown in the diagram). In our previous strategy, the most salient sibling of A111 (i.e., A112) would win the next attention if a positive answer is checked from the “view details” flag of A11. This strategy has advantages of simplicity and following the closest previous top-down setting to the higher level grouping (the parent A11 of A111). Here we present an improved strategy for such hierarchical attention shifts.

Suppose $S(X)$ represents the salience of any grouping X. Assume A and B are the most salient of the competitive groupings and $S(A) > S(B)$. Grouping A (or B) has a multi-level hierarchical structure. Then a tree-like data structure can be used to illustrate these structured groupings. Let the salience of the sub-groupings that have the same closest parent be decreasing from the left to the right. Let $A_{i_1,i_2,\ldots,i_j}$ be the current attended sub-grouping at the level j of A. When $i = 0$ or $j = 0$, $A_{i_1,i_2,\ldots,i_j} = A$. Thus the first level sub-groupings of A are $\ldots A_{i_1} A_{i_1+1} \ldots$, the first level sub-groupings of A_{i_1} are

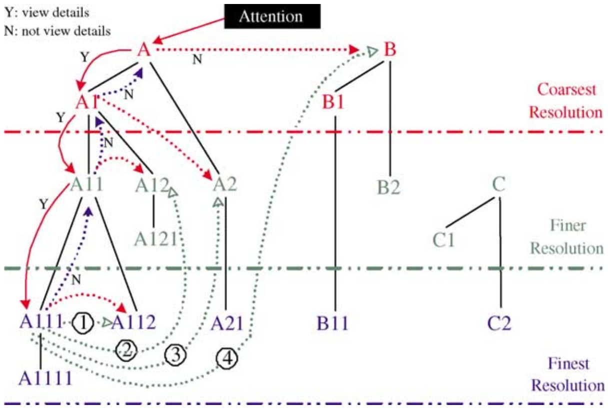

Fig. 25. Diagram of attentional movements in hierarchical selectivity operating on multi-level structured groupings. Red arrows: attentional movements. Blue arrows: feed-back checker of "view details" flag. Green arrows: possible winners competing for the next attention.

$\ldots A_{i_1,i_2} A_{i_1,i_2+1} \ldots$, and the rest is deduced by this analogy. Clearly, all sub-groupings left of $A_{i_1,i_2,\ldots,i_j}$ have already been attended or ignored. $A_{i_1,i_2,\ldots,i_j+1}$ is the most salient unattended sibling of the current attended grouping and $A_{i_1,i_2,\ldots,i_{j-1}+1}$ is the most salient unattended sibling of its parent. When attending $A_{i_1,i_2,\ldots,i_j}$, if a negative answer is given to the "view details" flag of top-down attentional setting or this sub-grouping has no child, the next potential winner to gain attention is produced by the following rules:

(1) if $A_{i_1,i_2,\ldots,i_j+1} = A$ then attention shifts to grouping B;
(2) otherwise attention shifts to the sub-grouping X with salience:

$$S(X) = MAX\big\{S(A_{i_1+1}), S(A_{i_1,i_2+1}), \ldots, S(A_{i_1,i_2,\ldots,i_{j-1}+1}), S(A_{i_1,i_2,\ldots,i_j+1})\big\}. \quad (23)$$

We applied this improved hierarchical selectivity to the natural scene shown in Fig. 13. Here the entire scene is re-segmented into seven top groupings, as shown in Fig. 26 (Graph B) by different colour lines. The identifiers of different groupings and their sub-groupings are also given in Graph B. Certain sub-groupings which are segmented within each top grouping are identified and the remainder (such as green grass in grouping 7 or trees in grouping 3) are denoted "others" in Graph A of Fig. 26. The "view details" flags of the parent groupings of the small white stripes in the road, trees in the lawns, and some tiny words (and symbols) below the "30" speed limit sign were answered "0" (positive) for the first attending (the first stripe, word or symbol) and "1" (negative) thereafter. Thus most sub-sub-groupings such as those within sub-groupings 6-1, 6-2, and 6-3 of top grouping 6 are also abbreviated as "others" in Graph A, except several first attended sub-sub-groupings (for example, grouping 6-1-1). The $1/\rho$ parameter of Gaussian distance is set to 25% for the global competition between the seven regions and 4% for the local competition within these regions.

Through the improved hierarchical selectivity, more natural attentional movements are clearly seen (Graph C in Fig. 26. Note here attention is assumed to shift to the center of

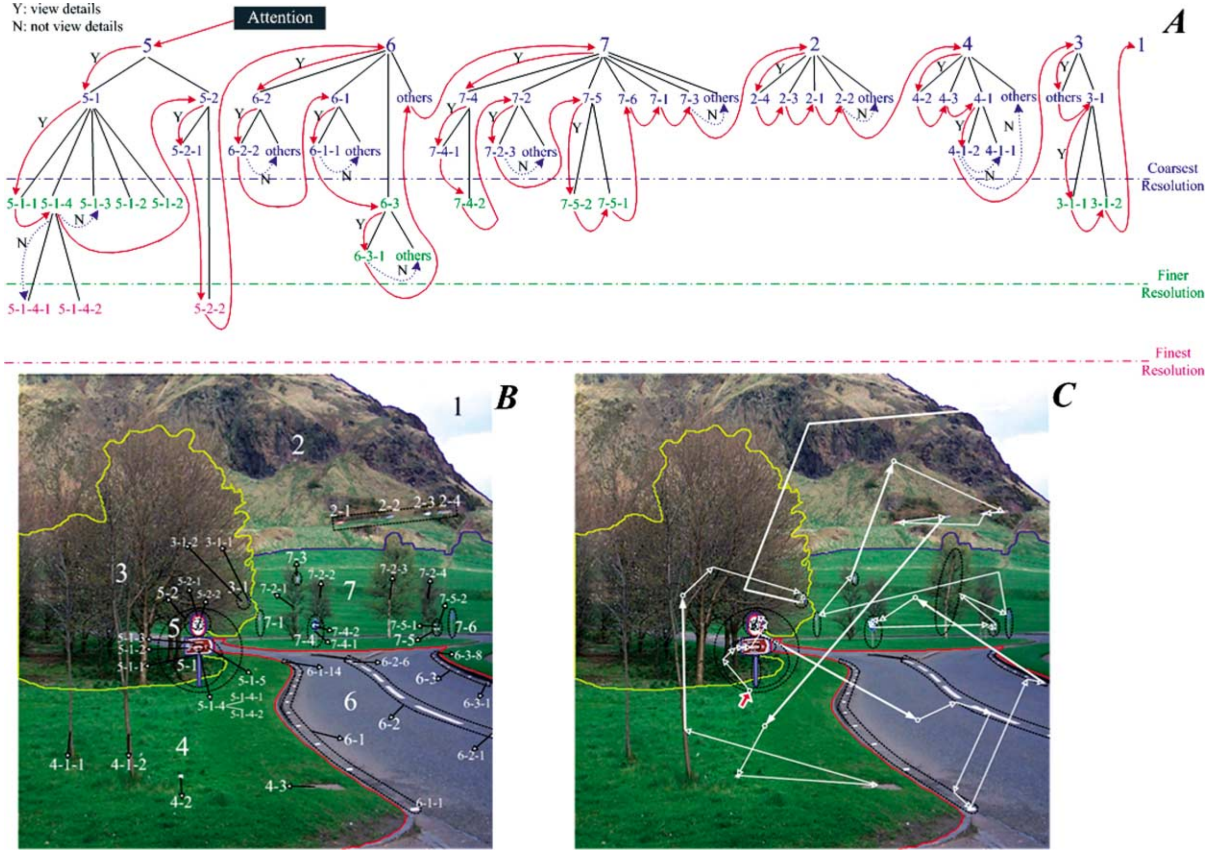

Fig. 26. The overall attentional movements on the natural scene produced by the improved strategy for hierarchical selectivity. Red arrows with a hollow circle indicate that attention goes to a top grouping and then shifts to the sub-groupings respectively. The dotted ellipses are not the sub-grouping boundaries and only used to conveniently show attention movements.

mass of the attended grouping). The complete hierarchical selectivity procedure for this scene is shown in Graph A in which the representations have the same meanings as those in Fig. 25.

4. Conclusion

The mechanisms of object-based and space-based visual attention have been widely investigated in psychophysics and neuroscience research, however, modelling visual attention in computer vision is a quickly growing field, especially for building computable models of covert attention. Until now, to our knowledge, although some computable models for space-based covert attention such as Koch and Itti's saliency-based attention model [45,53] have been successfully built, no computational model for object-based attention has been developed.

We have presented a computable model of hierarchical object-based attention for computer vision. It suggests that object-based and space-based attention can be integrated by using grouping-based salience to deal with dynamic visual tasks. By using the integrated competition of proto-objects based on groupings, the selectivity of attention by objects, locations and features can cooperatively work together. We demonstrated the behaviour of the model on a number of synthetic and real images. The experimental results showed that its performance concurs with the main findings in the psychophysical literature on object-based or space-based visual attention. Also, the model shows a good performance of selectivity by objects, by features, by spatial regions, and by their groupings on complex natural scenes. Such successful performances depend on three factors that we proposed here:

- grouping-based saliency evaluation;
- integrated competition between groupings;
- hierarchical selectivity.

With the grouping-based saliency mechanism, the pop-out of objects and their groupings can be evaluated in a uniform computational framework. By using hierarchical selectivity to drive attentional movements, the multiple selectivities of objects, features, regions, and their groupings in multiscale resolutions can be performed in an integrated selection architecture. To our knowledge, the model proposed in this paper is the first implemented model of object-based visual attention and of integrated object-based visual attention with space-based visual attention in computer vision.

However, there are still several limitations to the current model besides the above strengths. One limitation is that we have not yet built a satisfactory method to deal with the grouping processing. This is a great challenge not only for visual attention but also for computer vision. Another limitation is that we did not present here a complete theory of goal-driven effects on visual attention, which is necessary for understanding visual attention. Lastly, if we use a resolution-varying or retina-like operator at each attention movement, the model will simulate the attention behaviour of human eyes better, because human eyes have decreasing resolution from the fovea to the periphery of the retina. We are currently investigating these points.

Acknowledgements

The authors are grateful to Dr Fang Wang, Dr Laurent Itti, Dr Petko Faber, Dr Neil McCormick, and Dr Craig Robertson for their constructive suggestions.

References

[1] I. Ahrns, H. Neumann, Space-variant dynamic neural fields for visual attention, in: Proc. IEEE Computer Vision and Pattern Recognition, Fort Collins, CO, 1999, pp. 313–318.

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

[3] S. Baluja, D. Pomerleau, Expectation-based selective attention for visual monitoring and control of a robot vehicle, Robotics and Autonomous Systems 22 (1997) 329–344.

[4] M. Behrmann, R.S. Zemel, M.C. Mozer, Occlusion, symmetry, and object-based attention: Reply to Saiki (2000), J. Exp. Psych.: Hum. Percept. Perf. 26 (4) (2000) 1497–1505.

[5] C. Bundesen, A computational theory of visual attention, Phil. Trans. R. Soc. London B 353 (1998) 1271–1281.

[6] P. Burt, Attention mechanisms for vision in a dynamic world, in: Proc. Ninth International Conference on Pattern Recognition, Beijing, China, 1988, pp. 977–987.

[7] G. Carpenter, S. Grossberg, G. Lesher, The representation of visual salience in monkey parietal cortex, Nature 391 (1998) 481–484.

[8] J.J. Clark, N. Ferrier, Modal control of an attention vision system, in: Proc. IEEE Internat. Conf. Computer Vision, Tarpon Springs, FL, 1988, pp. 514–523.

[9] J.J. Clark, Spatial attention and latencies of saccadic eye movements, Vision Res. 39 (3) (1998) 583–600.

[10] V. Concepcion, H. Wechesler, Detection and localization of objects in time-varying imagery using attention, representation and memory pyramids, Pattern Recognition 29 (9) (1996) 1543–1557.

[11] W. Cowan, Evolving conceptions of memory storage, selective attention and their mutual constraints within the human information-processing system, Psychol. Bull. 104 (1988) 163–191.

[12] F. Crick, C. Koch, Towards a neurobiological theory of consciousness, Seminars in the Neurosciences 2 (1990) 263–275.

[13] http://www.dai.ed.ac.uk/CVonline.

[14] R. Desimone, J. Duncan, Neural mechanisms of selective visual attention, Ann. Rev. Neurosci. 18 (1995) 193–222.

[15] R. Desimone, Visual attention mediated by biased competition in extrastriate visual cortex, Phil. Trans. R. Soc. London B 353 (1998) 1245–1255.

[16] J. Driver, G.C. Baylis, Attention and visual object segmentation, in: R. Parasuraman (Ed.), The Attentive Brain, MIT Press, Cambridge, MA, 1998, pp. 299–325.

[17] J. Driver, G. Davis, C. Russell, M. Turatto, E. Freeman, Segmentation, attention and phenomenal visual objects, Cognition 80 (2001) 61–95.

[18] J. Duncan, Selective attention and the organization of visual information, J. Exp. Psychol. 113 (1984) 501–517.

[19] J. Duncan, G.W. Humphreys, Visual search and stimulus similarity, Psychological Rev. 96 (1989) 433–458.

[20] J. Duncan, Target and non-target grouping in visual search, Perception and Psychophysics 57 (1) (1995) 117–120.

[21] J. Duncan, Coordinated brain systems in selective perception and action, in: T. Iaui, J.L. McClelland (Eds.), Attention and Performance XVI, MIT Press, Cambridge, MA, 1996, pp. 549–578.

[22] J. Duncan, et al., Integrated mechanisms of selective attention, Curr. Opin. Biol. 7 (1997) 255–261.

[23] J. Duncan, Converging levels of analysis in the cognitive neuroscience of visual attention, Phil. Trans. R. Soc. London B 353 (1998) 1307–1317.

[24] H.E. Egeth, S. Yantis, Visual attention: Control, representation, and time course, Ann. Rev. Psychol. 48 (1997) 269–297.

[25] R. Egly, et al., Shifting visual attention between object and locations: Evidence from normal and parietal lesion subjects, J. Exp. Psychol. Hum. Percept. 123 (1994) 161–177.

[26] S. Engle, X. Zhang, B.A. Wandell, Colour tuning in human visual cortex measured with functional magnetic resonance imaging, Nature 388 (6637) (1997) 68–71.

[27] C.W. Eriksen, Y.Y. Yeh, Allocation of attention in the visual field, J. Exp. Psychol.: Hum. Percept. Perf. 11 (5) (1985) 583–597.

[28] C.W. Eriksen, J.D.St. James, Visual attention within and around the field of focal attention: A zoom lens model, Perception and Psychophysics 40 (4) (1986) 225–240.

[29] S. Exel, L. Pessoa, Attention visual recognition, International Conference on Pattern Recognition, Brisbane, Australia, 1998.

[30] M.J. Farah, et al., “What” and “where” in visual attention: Evidence from the neglect syndrome, in: Unilateral Neglect: Clinical and Experimental, 1993, pp. 123–138.

[31] V. Ferrara, S. Lisberger, Attention and target selection for smooth pursuit eye movements, J. Neurosci. 15 (11) (1995) 7472–7484.

[32] G.R. Fink, et al., Space-based and object-based visual attention: Shared and specific neural domains, Brain 120 (1997) 2013–2028.

[33] C.H. Folk, W.R. Remington, J.H. Wright, The structure of attentional control: contingent attentional capture by apparent motion, abrupt onset, and color, J. Exp. Psychol.: Hum. Percept. Perf. 20 (2) (1994) 317–329.

[34] J.P. Gottlieb, et al., The representation of visual salience in monkey parietal cortex, Nature 391 (6666) (1998) 481–484.

[35] H. Greenspan, S. Belongie, R. Goodman, P. Persona, S. Rakshit, C.H. Anderson, Overcomplete steerable pyramid filters and rotation invariance, in: Proc. IEEE Computer Vision and Pattern Recognition, Seattle, WA, 1994, pp. 222–228.

[36] W.E.L. Grimson, et al., An active visual attention system to play “Where’s Waldo”, in: Proc. Conference on Computer Vision and Pattern Recognition, Seattle, WA, 1994, pp. 85–90.

[37] S. Grossberg, et al., A neural theory of attentive visual search: interactions of boundary, surface, spatial and object representations, Psychological Rev. 10 (3) (1994) 470–489.

[38] S. Grossberg, How does the cerebral cortex work? Learning, attention, and grouping by the laminar circuits of visual cortex, Spatial Vision 12 (2) (1999) 13–185.

[39] S. Grossberg, R. Raizada, Contrast-sensitive perceptual grouping and object-based attention in the laminar circuits of primary visual cortex, Vision Res. 40 (2000) 1413–1432.

[40] T.D. Grove, R.B. Fisher, Attention in iconic object matching, in: Proc. BMVC96, Edinburgh, 1996, pp. 293–302.

[41] D. Heinke, G.W. Humphreys, SAIM: A model of visual attention and neglect, in: Proc. International Conference on Artificial Neural Networks, New York, 1997, pp. 913–918.

[42] G.W. Humphreys, SEarch via recursive rejection (SERR): A connectionist model of visual search, Cognitive Psychology 25 (1993) 43–110.

[43] G.W. Humphreys, Neural representation of objects in space: A dual coding account, Phil. Trans. R. Soc. London B 353 (1998) 1341–1351.

[44] J.E. Hoffman, Visual attention and eye movements, in: H. Pashler (Ed.), Attention, Psychology Press, 1998, pp. 119–154.

[45] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Machine Intel. 20 (11) (1998) 1254–1259.

[46] L. Itti, C. Koch, A saliency-based search mechanism for overt and covert shifts of visual attention, Vision Res. 40 (10–12) (2000) 1489–1506.

[47] E.N. Johnson, M.J. Hawken, R. Shapley, The spatial transformation of color in the primary visual cortex of the macaque monkey, Nature 4 (4) (2001) 409–416.

[48] D. Kahneman, A. Henik, Perceptual organization and attention, in: M. Kubovy, J.R. Pomerantz (Eds.), Perceptual Organization, Erdbaum, Hillsdale, NJ, 1984, pp. 181–211.

[49] P. Kaiser, R.M. Boynton, Human Color Vision, 2nd Edition, Optical Society of America, 1996.

[50] S. Kastner, L.G. Ungerleider, Mechanisms of visual attention in the human cortex, Ann. Rev. Neurosci. 23 (2000) 315–341.

[51] Y.B. Kazanovich, R.M. Borisyuk, Dynamics of neural networks with a central element, Neural Networks 12 (1999) 441–454.

[52] E. Kowler, et al., The role of attention in the programming of saccades, Vision Res. 35 (13) (1995) 1897–1916.

[53] C. Koch, S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry, Human Neurobiology 4 (1985) 219–227.

[54] A.F. Kramer, A. Jacobson, Perceptual organization and focused attention: The role of objects and proximity in visual processing, Perception and Psychophysics 50 (1991) 267–284.

[55] V.I. Kryukov, An attention model based on the principle of dominanta, in: A.V. Holden, V.I. Kryukov (Eds.), Neurocomputers and Attention I: Neurobiology, Synchronization and Chaos, Manchester University Press, Manchester, 1991, p. 319.

[56] D. LaBerge, Attentional Processing: The Brain's Art of Mindfulness, Harvard University Press, Harvard, 1995.

[57] N. Lavie, Perceptual load as a necessary condition for selective attention, J. Exp. Psychol.: Hum. Percept. Perf. 21 (1995) 451–468.

[58] N. Lavie, J. Driver, On the spatial extent of attention in object-based selection, Perception and Psychophysics 58 (1996) 1238–1251.

[59] G.D. Logan, The CODE theory of visual attention: An integration of space-based and object-based attention, Psychological Rev. 103 (4) (1996) 603–649.

[60] S.J. Luck, Neurophysiology of selective attention, in: H. Pashler (Ed.), Attention, Psychology Press, 1998, pp. 257–295.

[61] R.M. McPeek, et al., Saccades require focal attention and are facilitated by a short-term memory system, Vision Res. 39 (1999) 1555–1566.

[62] A. Nemcsics, Color Dynamics, Akademiai Kiad, Budapest, 1993.

[63] E. Niebur, et al., An oscillation based model for the neuronal basis of attention, Vision Res. 33 (1993) 2789–2802.

[64] E. Niebur, C. Koch, A model for the neuronal implementation of selective visual attention based on temporal correlation among neurons, J. Neurosci. 1 (1994) 141–158.

[65] H.C. Nothdurft, The conspicuousness of orientation and motion contrast, Spatial Vision 7 (4) (1993) 341–363.

[66] B.A. Olshausen, et al., A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information, J. Neurosci. 13 (11) (1993) 4700–4719.

[67] S.E. Palmer, Vision Science-Photons to Phenomenology, MIT Press, Cambridge, MA, 1999.

[68] H. Pashler, The Psychology of Attention, MIT Press, Cambridge, MA, 1998.

[69] G.A. Patel, K. Sathian, Visual search: bottom-up or top-down?, Frontiers in Bioscience 5 (2001) 169–193.

[70] M.E. Posner, Orienting of attention, Q. J. Exp. Psychol. 32 (1980) 3–25.

[71] E.O. Postma, et al., SCAN: A scalable model of attentional selection, Neural Networks 10 (1997) 993–1015.

[72] C. Poynton, Frequently asked questions about color, http://www.inforamp.net/~Poynton/.

[73] A.L. Ratan, The role of fixation and visual attention in object recognition, MIT AI-TR-1529, July, 1995.

[74] D.J. Robinson, S.E. Peterson, The pulvinar and visual salience, Trends in Neuroscience 15 (4) (1992) 127–132.

[75] I.A. Rybak, et al., A model of attention-guided visual perception and recognition, Vision Res. 38 (1998) 2387–2400.

[76] B.J. Scholl, Objects and attention: The state of the art, Cognition 80 (2001) 1–46.

[77] A. Shokoufandeh, et al., View-based object recognition using saliency maps, Image and Computing 17 (1999) 445–460.

[78] A.M. Sillito, et al., Visual cortex mechanisms detecting focal orientation discontinuities, Nature 378 (1995) 492–496.

[79] W. Singer, C.W. Gray, Visual feature integration and the temporal correlation hypothesis, Ann. Rev. Neurosci. 18 (1995) 555–586.

[80] G. Sela, M.D. Levine, Real-time attention from robotic vision, Real-Time Imaging 3 (1997) 173–194.

[81] Y. Sun, Object-based visual attention and attention-driven saccadic eye movements for machine vision, PhD Thesis, the University of Edinburgh, 2003.

[82] B. Takacs, H. Wechsler, A dynamic and multiresolution model of visual attention and its application to facial landmark detection, Computer Vision and Image Understanding 70 (1) (1998) 63–73.

[83] A. Treisman, G. Gelade, A feature integration theory of attention, Cognition Psychology 12 (1980) 97–136.

[84] A. Treisman, Features and objects: The fourteenth Bartlett Memorial lecture, Q. J. Experimental Psychology 40A (1988) 201–237.

[85] A. Treisman, The perception of features and objects, in: A. Baddeley, L. Weiskrantz (Eds.), Attention: Selection, Awareness, and Control, Uarendon Press, Oxford, 1993, pp. 5–35.

[86] J.K. Tsotsos, et al., Modelling visual attention via selective tuning, Artificial Intelligence 78 (1995) 507–545.

[87] M. Usher, N. Donnelly, Visual synchrony affects binding and segmentation in perception, Nature 394 (1998) 179–182.

[88] E.J. Chichilnisky, B.A. Wandell, Trichromatic opponent color classification, Vision Res. 39 (20) (1999) 3444–3458.

[89] B.A. Wandell, Computational neuroimaging: color representations and processing, in: M.S. Gazzaniga (Ed.), New Cognitive Neuroscience, MIT Press, Cambridge, MA, 1999.

[90] C.F. Westin, et al., Attention control for robot vision, in: Proc. IEEE Computer Vision and Pattern Recognition, San Francisco, CA, 1996, pp. 18–20.

[91] J.W. Wolfe, Guided Search 2.0: A revised model of visual search, Psychonomic Bulletin and Review 1 (1994) 202–238.

[92] J.W. Wolfe, Visual search, in: H. Pashler (Ed.), Attention, Psychology Press, 1998, pp. 13–73.

[93] G. Wyszechi, W.S. Stiles, G. Wyszecki, G. Wyszecki, Color Science: Concepts and Methods, Quantitative Data and Formulae, 2nd Edition, Wiley, New York, 2000.

[94] S. Yantis, Control of visual attention, in: H. Pashler (Ed.), Attention, Psychology Press, 1998, pp. 223–256.