International Journal of Neural Systems, Vol. 17, No. 4 (2007) 1–14

World Scientific
www.worldscientific.com

ATTENTION AND VISUAL SEARCH

ANTONIO J. RODRIGUEZ-SANCHEZ*, EVGUENI SIMINE†
and JOHN K. TSOTSOS‡
Centre for Vision Research and Department of Computer Science and Engineering,
York University, 4700 Keele St., Toronto, ON M3J1P3, Canada
**ajrs@cse.yorku.ca*
**www.cse.yorku.ca/∼ajrs*
†*eugene@cse.yorku.ca*
‡*tsotsos@cse.yorku.ca*
‡*www.cse.yorku.ca/∼tsotsos*

Selective Tuning (ST) presents a framework for modeling attention and in this work we show how it performs in covert visual search tasks by comparing its performance to human performance. Two implementations of ST have been developed. The Object Recognition Model recognizes and attends to simple objects formed by the conjunction of various features and the Motion Model recognizes and attends to motion patterns. The validity of the Object Recognition Model was first tested by successfully duplicating the results of Nagy and Sanchez. A second experiment was aimed at an evaluation of the model's performance against the observed continuum of search slopes for feature-conjunction searches of varying difficulty. The Motion Model was tested against two experiments dealing with searches in the visual motion domain. A simple odd-man-out search for counter-clockwise rotating octagons among identical clockwise rotating octagons produced linear increase in search time with the increase of set size. The second experiment was similar to one described by Thorton and Gilden. The results from both implementations agreed with the psychophysical data from the simulated experiments. We conclude that ST provides a valid explanatory mechanism for human covert visual search performance, an explanation going far beyond the conventional saliency map based explanations.

Keywords: Visual search; attention; features; conjunction; motion; object recognition.

1. Introduction

The breadth of functionality associated with attentional processing can easily be seen in several overviews (e.g. Refs. 1, 2). One of the most studied topics and with a very significant literature is that of visual search. Visual search experiments formed the basis and motivation for the earliest of the influential models (e.g. Refs. 3, 4). Yet, no satisfactory explanation of how the network of neurons that comprise the visual cortex performs this task exists. Certainly, no computational explanation or model exists either.

In a visual search experiment, the task is to look for a target defined by one or more features among a set of distractors that are different from the target but may share one or more features with it. When target and distractors are the same except for one feature, it is commonly known as feature visual search and the target seems to pop-out (e.g. a red vertical bar among a set of green vertical bars). When there are two different kinds of distractors and the target shares a feature with each one of the two types of distractors, this search is referred as conjunction search and it requires more time to find the target (e.g. Look for a red vertical bar among red horizontal bars and green vertical ones). Decades of psychophysical experimentation have analyzed response-time (RT) as a function of the number of distractors for most of the different features under thousands of different situations.[5]

The analysis of RT × size slopes has been widely used to propose different theories on how the brain works for such tasks. One of the most influential was the Feature Integration Theory,[3] which proposed that feature search was the result of a parallel process while a conjunction search was the result of a serial search. More recent models[6,7] have rejected that hypothesis, proposing a visual search continuum directly related to the similarity among target and distractors.

Some recent models of attention[8,9] have been compared to human eye movement tracks -overt attention- as validation; but this is not the same as visual search data which is almost exclusively covert, with no eye movement. Visual attention involves much more than simply the selection of next location to fixate the eyes or camera system, regardless of the fact that the vast majority of all computational approaches to attention focus on this issue exclusively. That humans are able to attend to different locations in their visual field without eye movements has been known since Ref. 8. Further, eye movements require a shift of visual attention to precede them to their goal, Ref. 1 surveys relevant experimental work).

Attentional models have matured sufficiently so that this broader problem of attention can now be confronted. This paper makes several steps towards the development of such an explanation expanding the Selective Tuning model[11,12] and comparing performance with existing visual search psychophysical performance. This is done with simple colored shape stimuli as well as with motion stimuli.

The rest of the paper is organized as follows: Section 2 describes the Object Recognition model with its two main pathways, shape analysis and color analysis, and tests the model with visual search experiments. Section 3 deals with testing the Motion Model in the case of moving targets from a set of distractors in different conditions. We finally present our conclusions in Sec. 4. The results of both implementations of the Selective Tuning[11,12] model is compared with psychophysical experiments extracted from the literature, obtaining comparable results to those.

2. Object Recognition Model

Given a scene with several objects, the model's purpose is to find a particular object that has been presented previously.

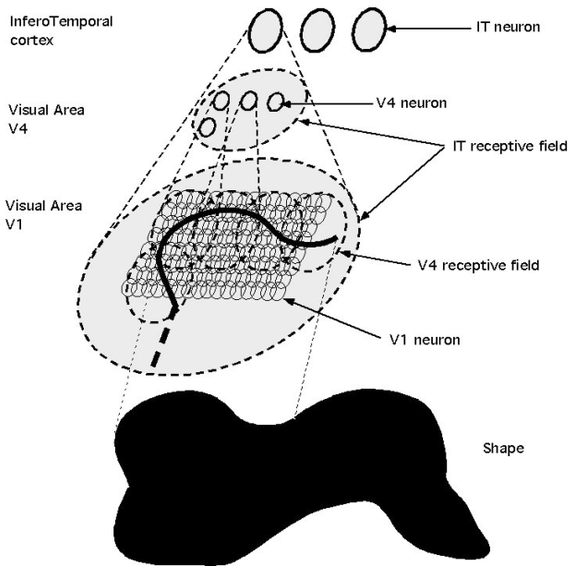


Fig. 1. Architecture of shape pathway.

The model structure is a two-pathway pyramid with information flowing from the input to the top of the pyramid and from the top to the bottom providing feedback. Each one of the two pathways analyze the visual input in a different way, one extracts color information, while the other extracts information about the shape of the objects (Fig. 1).

The model mimics the human visual pathway for object recognition, simulating four visual areas: LGN, V1, V4 and IT. Each area is organized into feature maps and each feature map encodes the visual field in a unique way. The model comprises a total of 22 feature maps.

Information first flows from the input to area LGN and V1. LGN extracts three color feature maps (red, green and blue). V1 is composed of edge detectors organized in 8 feature planes (each containing neurons tuned to one of 8 directions). Two additional feature maps in V1 compute center-surround color differences from the LGN color feature maps. Information from V1 flows to V4, which comprises 8 feature maps for curvature. Finally, IT neurons encode a representation of the whole object based on curvature and color differences.

Our strategy follows the sequence of events in a human visual search experiment, that is, a subject is first shown the target on a blank display, then is

shown the test display to be searched. Similarly, the system is first shown the target and extracts a representation of it. This representation is used to bias the subsequent search when the test display is presented. When the test display is presented, biased shape and color analysis proceed in parallel in a feed-forward manner, then the Selective Tuning[11] feedback attentive process is applied. The different stages of processing are explained in more detail in the following sections.

2.1. *Shape analysis*

The shape processing pathway (Fig. 1) is inspired by Pasupahy and Connor.[13] Visual Area V1 contains neurons that perform edge analysis. Gabor filters[14] are used with 8 different orientations:

$$\begin{aligned} G(x,y) &= e^{-(a^2x'+b^2y')} \times e^{-(j2pfx')} \\ x' &= xcos(q) + ysin(q) \\ y' &= -xsin(q) + ycos(q) \end{aligned} \tag{1}$$

where α and β are the sharpness of the Gaussian major and minor axis, with values of 1 and 0.25 in our case; f is the frequency and θ is the orientation.

The size of the neuron's receptive field is 16×16 pixels. The output of V1 neurons is 8 feature planes, representing edges at 8 orientations. Non-maximal suppression[15] is applied in order to reduce the Gabor filter output to a 1-2 pixel wide images as a preprocessing for the next visual layer. The output from V1 neurons feeds into V4.

V4 neurons compute curvature values based on orientation changes from groups of adjacent V1 neurons. For example, if a V1 neuron in a V4 receptive field had its highest response for $\theta = 0$ and another adjacent one had a high response for $\theta = \pi/4$, we would have a corner. If both orientations were equal, it would correspond to a straight line.

Curvature for V4 is then defined as:

$$\begin{gathered} curv = min(|\theta_1 - \theta_2|, 2\pi - |\theta_1 - \theta_2|) \\ curv \in [0, \pi) \end{gathered} \tag{2}$$

where θ_1 and θ_2 are the orientations of two V1 cells. A value of π can be added to θ_1 and/or θ_2 depending on the neurons' relative positions inside the V4 receptive field due to the fact that the same Gabor filter orientation can account for two different angles. The activation value of the V4 neuron is the summed activations from the V1 neurons used to obtain the curvature. V4 neurons receptive field comprise groups of 4×4 V1 neurons.

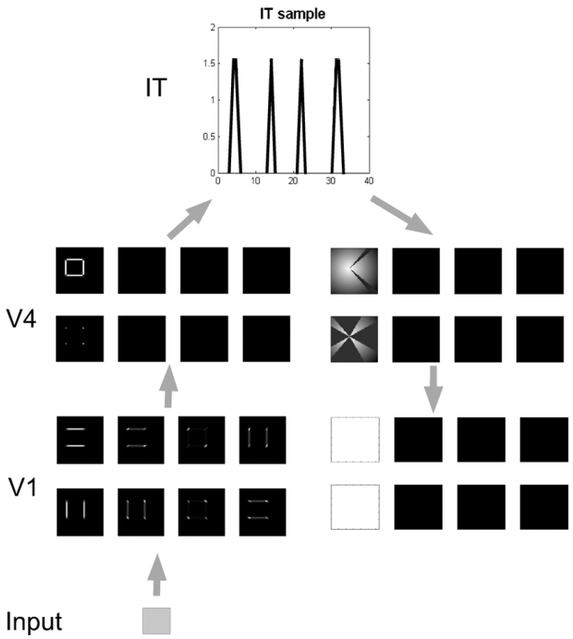


Fig. 2. Shape analysis on target stimulus. Left (bottom-up): Edges are extracted in V1 at each different orientation, then in V4 curvatures are calculated, finally IT computes the curvature×position representation[13] Right (top-down bias): From such a representation, in V4 feature planes that do not have values of curvature corresponding to the object are inhibited (black), and in these V4 feature planes, neurons that are not at the proper location are inhibited as well. In V1, neurons that do not contribute to those V4 feature planes are also inhibited, only allowing the Gabor filters corresponding to the orientations that feed into the non-inhibited V4 neurons.

V4 neurons' output is 8 2D feature maps that encode for the difference of curvature among groups of V1 neurons. This output feeds into IT at the very top of the hierarchy (Fig. 2). The receptive fields of IT neurons comprise an area of 32×32 V4 neurons (that is, 128×128 pixels). The center of mass is calculated for every group of V4 neurons as the mean of the V4 neuron coordinates where responses are different from zero. Then, at each angular position (in 10 deg bins), its curvature is computed in agreement with the representation proposed by Ref. 13, obtaining a histogram-like representation for IT neurons where one axis correspond to the angular position

(λ) and the other coordinate is the curvature *curv* for that position (Fig. 2):

$$\lambda = round\left[\frac{\tan^{-1}\left(\frac{y - centroid_y}{x - centroid_x}\right)*18}{\pi}\right] \qquad IT(\lambda) = curv \tag{3}$$

The term $18/\pi$ is for the angular position to be in 10 deg bins.

All neuron relative sizes were chosen to correspond closely to the neurophysiological measured sizes[16] considering a distance of 30 cm (usual psychophysical distance) to a 1280×1024 display. Neurons' receptive fields are overlapped.

2.2. *Color analysis*

The processing of color follows a centre-surround analysis.[17] A first layer (LGN) extract 3 feature maps for red (R), green (G) and blue (B) responses. In the upper layer (V1), surround values for red-green (RG), green-red (RG), blue-yellow (BY) and yellow-blue (YB) are extracted following most models (e.g Ref. 8):

$$RG = \frac{(R - G)}{Luminance} \qquad BY = \frac{(B - Y)}{Luminance} \tag{4}$$

RG feature plane also accounts for GR differences, the same applies to the BY feature plane.

As in the Shape analysis, color neurons at every level of the hierarchy are also inhibited if they do not share the values corresponding to center-surround and color activations of the target.

2.3. *The bias stage*

After the representation of the object shape is obtained, V4 and V1 layers are biased (Fig. 2 right). In V4 only neurons that are not in the proper angular position and in the desired curvature feature planes are completely inhibited. For those V4 neurons not completely inhibited, a partial inhibition will be applied to those ones that are further apart to the object's center of mass, inhibition in this case is linearly proportional to the Euclidean distance to the object's center. At a lower level, the neurons inhibited in V1 correspond to those whose orientation values were related to the curvature inhibited V4 neurons.

2.4. *Recognition*

Before the presentation of the test display, the network is biased to expect the target stimulus. The point of this bias is to speed up search; it has been shown that advance knowledge of the target indeed speeds up detection in a test display.[18–20] However, erroneous knowledge of the target slows down overall search.[21] The processing is first biased by the presented object or target representation at the different visual layers of the network so that after the first feed-forward pass of processing the test display only locations with the desired target features will be considered. Then, the search begins after a feed-forward activation by considering the best matching IT neuron from the possible candidates containing non-biased features.

To determine how close is the shape to the desired shape, distance to the target IT histogram is computed, for this distance we used cummulative distance. This distance is very common for computing distances between histograms and it is used here due to the similar representation of IT neurons to a histogram:

$$d(p,q) = \sqrt{\sum_{i=0}^{L-1}\left(\sum_{u=0}^{i} p_u - \sum_{u=0}^{i} q_u\right)^2} \tag{5}$$

The activation of the neuron is inversely proportional to d. Both activation values for color and shape $\in [0, 1]$ and the activation of the candidate IT neuron is the addition of both values. Even though the object can be in the receptive field of the highest activated IT neuron, due to its large receptive field and even after the bias, it can accommodate other objects (that may even disturb the firing values of the IT neuron). Information is further filtered in the lower layers (V4, V1) by computing winner-take-all in a hierarchical fashion.[11] The WTA processes in V4 are grouped by curvature angle. There is a separate WTA process for each 10 deg bin (as determined by Eq. 2), i.e. a V4 neuron will only compete with neurons in the same bin. In V1 only those neurons connected with the V4 winners are considered, and the same process is applied when going from V1 to the image, finding the contour of the candidate object. Figure 3 shows an example of this process. Inhibition of return was implemented in by blanking the part of the input image corresponding to the analyzed object.

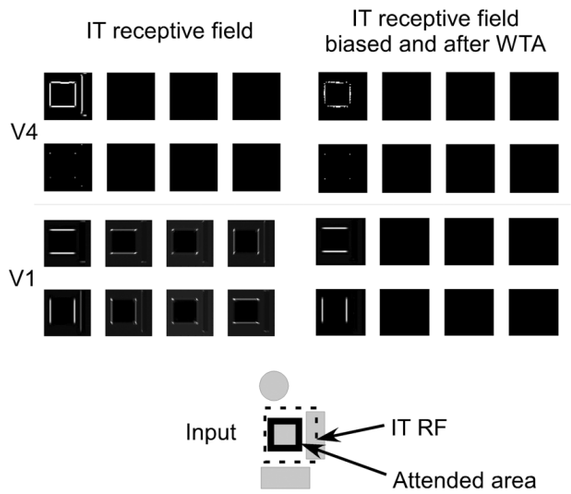


Fig. 3. Analysis of a scene (bottom): find the square. Left: V1 layer extracts edges, V4 neurons compute curvature. Here, inside the IT neuron receptive field (RF) lays the square and part of an object of no interest (rectangle). Right: V1 and V4 layers after attention. Layers in the hierarchy are first biased and information is later filtered through a winner take all process (See Ref. 11 for a full explanation). Thanks to this process, information is filtered such that the object of interest (square) is the only object that remains inside the IT neuron RF.

2.5. *Results*

We tested the model's behavior for different visual search conditions. For these tests, we followed different psychophysical experiments and we compared the results obtained from those works with the results obtained from the model.

But before testing the model for visual search, we performed a study on how the shape representation works for a simple recognition of silhouette-kind of objects.

2.5.1. *Silhouette search*

Pasupathy and Connor[13] used simple icons to infer how neurons responded to shapes. As a result, we first tested the shape analysis component of our model with 2D silhouettes.

To test the model the silhouette database from Ref. 22 was used. The architecture was fed with different silhouettes of objects, animals, cars, planes, etc. Then, scenes were constructed with such element silhouettes and the response corresponding to the scene IT neurons were evaluated. The IT neuron from the scene with a closer response (in terms of distance) to the neurons representation in the database was recognized as containing the object represented.

The test scene images were 512×512 pixels. IT neuron's receptive field were 128×128 as described previously and there was an IT neuron every ten pixels starting from the coordinates (64, 64) until coordinates (448, 448).

Figures 4 and 5 show different IT neurons and the object corresponding to their highest response (inside a dashed-line box). We show how the system performs when the whole object is present (Fig. 4) and in conditions when the objects are partially presented (Fig. 5). The system works pretty well in both cases. Although the silhouette is usually at the center of the neuron when training, the winning IT neuron doesn't need to have the object exactly at its center, but we can see that this is usually the case in accordance to Ref. 13. Figure 5a shows how the system behaves when there is partial information about the target objects in the scene. We can see that the model finds correctly every object, even when information is quite incomplete (e.g. the plane). Note that in these cases the IT receptive field center is not so

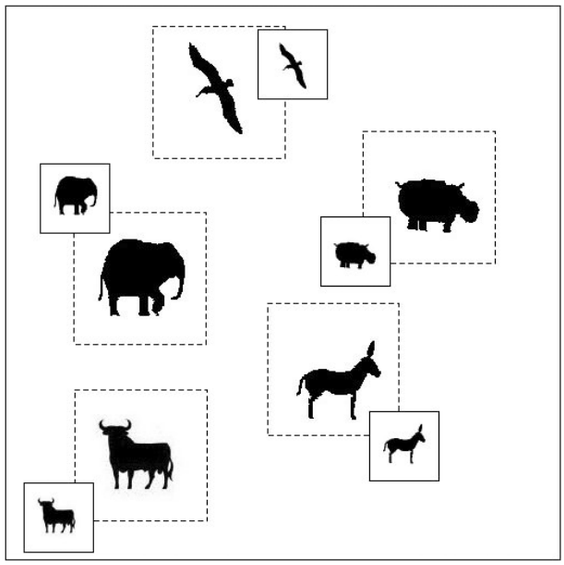
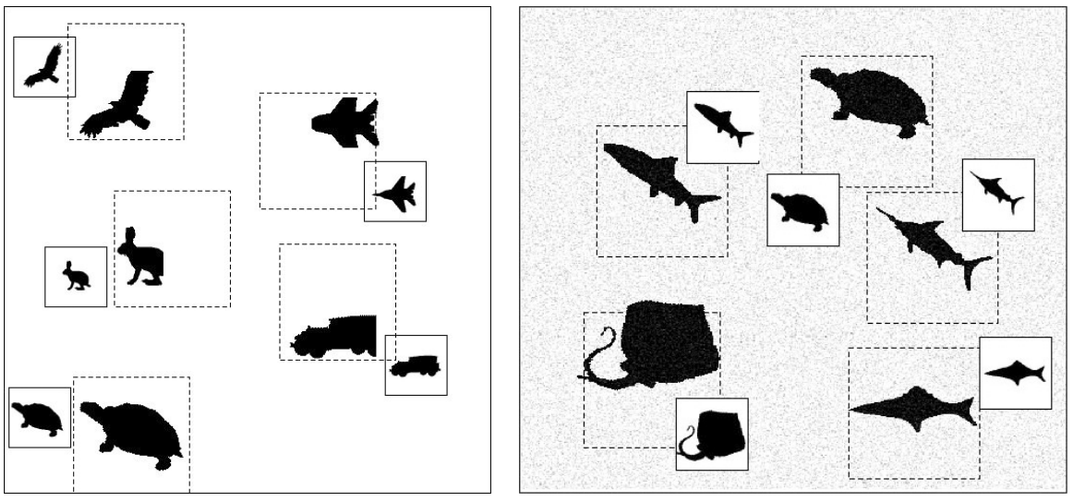


Fig. 4. Example of silhouette recognition. Recognition of complete objects. The dashed-line box correspond to the winning IT neurons receptive with fields that correspond to the silhouette being look for (shown in the small continuous-line box close to it).



(a) (b)

Fig. 5. Examples of silhouette recognition. (a) Recognition of partially presented silhouettes. (b) Recognition of scenes with gaussian noise. The dashed-line box correspond to the winning IT neurons receptive with fields that correspond to the silhouette being look for (shown in the small continuous-line box close to it).

close to the object's center, while if it is in its full shape that is usually the case (turtle).

In Fig. 5b we show how the representation performs for a case with Gaussian noise ($\mu = 0$, $\sigma^2 = 0.01$). The representations shows to be quite robust to noise. The only case where the winning neuron is not the optimal, corresponds to the ray, but a neuron very close to it corresponds to the winning neuron.

2.5.2. *Efficiency in visual search*

Recently, it has been shown that conjunction searches (See Ref. 5 for a review) may exhibit shallower slopes than those found by Ref. 3, and there seems to exist a continuum from efficient to inefficient visual search. An interesting theory is the one proposed by Ref. 6, they were the first to argue that visual search is influenced by the similarity between target and distractors, they stated that visual search is harder when target and distractors are more similar, but it is easier when this similarity decreases, this theory have been supported by later experiments.[23,24]

For so we decided to test the model first with an experiment concerning the similarity hypothesis. One that studies a fundamental basic feature (color) and known in the psychophysical community is Ref. 23, this is the experiment we first replicate.

In our second experiment we test the search continuum and we compare the performance of the model for feature search, conjunction search and inefficient search.

Feature search is a search where a target is distinguished from the distractors by a single feature such as color, shape or orientation. In our second experiment we will use the term feature search to refer to a classical psychophysical feature search experiment defined by its efficiency and for so, having a big dissimilarity between target and distractors. As shown in experiment 1, that feature search is efficient is not always the case.

In conjunction search a target is defined by a conjunction of two features. Finally, following Ref. 5, we use the term inefficient search for those visual search experiments that are more difficult than the classical conjunction search. Let's note that, strictly speaking, inefficient search is also a conjunction search, but we will use a different notation to distinguish it from the classical conjunction search.

The sample was given as input in a 128×128 pixel image, and the scenes were 640×640 pixels. In our first experiment we will test a known feature: color, and how the model performs under two different similarities of colored objects.

Summarizing, we first follow a known study[23] about color similarities and compare our results with

those of such study. In a second experiment we study more deeply the continuum efficient-inefficient search with Selective Tuning. We follow three known experiments and as before, compare our results with those.

Experiment 1: Color differences

Method: In this experiment we study how the model performs in a color similarity search. We try here to simulate experiment from Ref. 23, who showed that feature search can be inefficient if the differences in color are small. We used the CIE values from their experiments converted to RGB with a fixed luminance (Y) of 0.25. The task is to find the redder circle among 5, 10, 15, 20 and 25 distractors for two conditions: small and large color differences. The target and distractors were randomly positioned on a black background. The least mean squares method was used to fit the straight line into the set of points.

Results: An example is shown in Fig. 6 left, where, when there are small differences between the target and the distractors, a much larger number of attentional shifts are needed to find the target. Fig. 6 right shows how the number of attentional shifts increases as the set size increases. This experiment reports similar results to Ref. 23 where color search is inefficient if color difference is small between target and distractors (slope = 0.39) and efficient if the difference is large (slope = 0.01).

Experiment 2: Feature, conjunction and inefficient search

Bichot and Schall[25] showed that monkey visual search reaction times are comparable to human, namely they show that the conjunction of two different features (shape and color) is steeper than feature search, but shallower than what was obtained by Ref. 3. They report slopes of 3.9 ms/item. Searching for a rotated T among rotated Ls, Ref. 26 reported that this search was quite inefficient (20 msec/item), and less efficient than conjunction searches. To find a T among Ls is more inefficient than a conjunction search, which is less efficient than a simple feature search.

Method: In this experiment we study how the model performs in a simple feature search, a conjunction search and an inefficient search. Conjunction search was similar to that of Ref. 25. The stimuli were crosses and circles, red or green colored. The task was to find a red circle among green circles and red crosses, here we used 8, 12, 16, 18, 22 and 24 distractors. Feature search was a simplification of the previous conjunction search, that is, to look for a circle among crosses. For inefficient search, a rotated T was to be found among Ls rotated at 0, 90 and 180 degrees, in this case we used 6, 9, 12, 15, 18 and 21 distractors. Analysis was the same as for previous experiments.

Results: An example of searching for a T among Ls is shown in Fig. 7 many atttentional shifts are

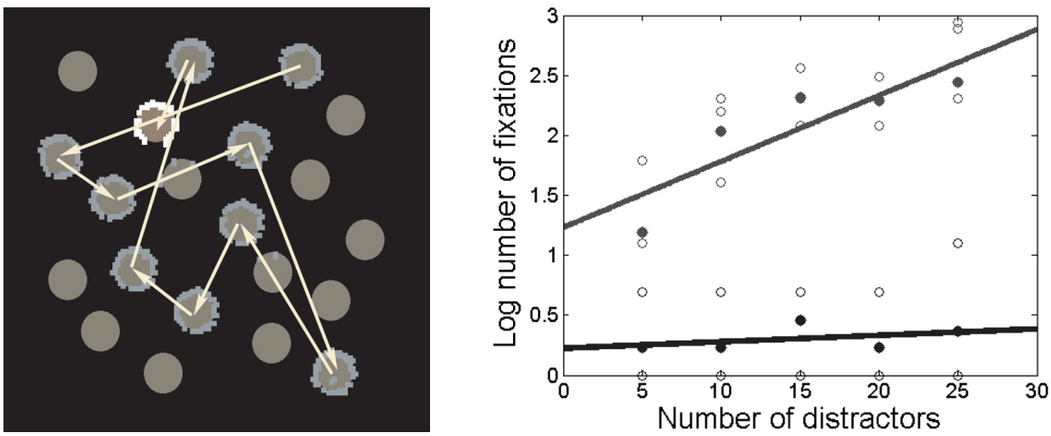


Fig. 6. Visual Search Results. Left: Example where the target and distractors have small color differences, 10 shifts of attention were needed to find the redder item (white outline). Right: The number of fixations as a function of set size. Gray line: large color difference, black line: small color difference.

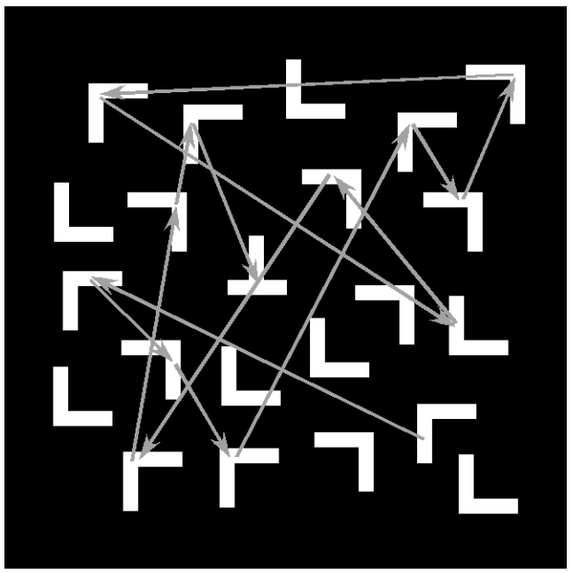


Fig. 7. Inefficient search: Find the rotated T among 21 *L*s, 14 fixations were needed to find the *T*.

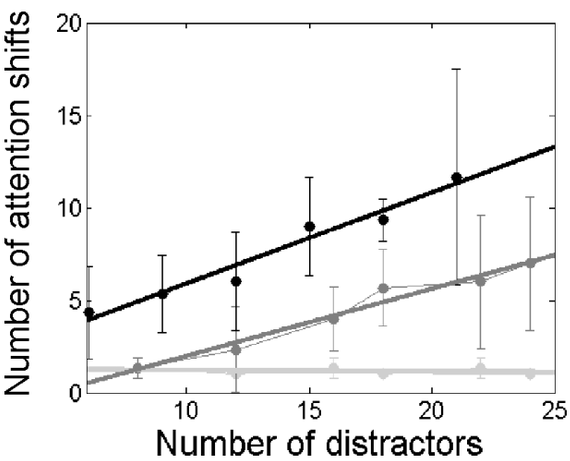


Fig. 8. The number of shifts of attention as a function of set size for feature search (light gray), conjunction search (gray) and inefficient search (black).

needed to find the target. Figure 8 shows the number of attentional shifts as the set size increases for the feature search (find a circle among arrows), conjunction search (find a red circle among red arrows and green circles) and inefficient search (find a rotated *T* among *L*s). The figure shows how the steepest fitted line is the one corresponding to looking for a *T* among *L*s (inefficient search, slope of 0.49) experiment, followed by conjunction search (slope of 0.36) and feature search is practically flat (slope of 0.00). These results are in accordance with the continuum from efficient to inefficient search psychophysical experiments have shown (see 5 for a review).

Discussion

The above results show the ability of the Object Recognition Model to perform visual search. The reaction time is shown based on the number of attentional shifts. We performed easy feature search, difficult feature search, conjunction search and inefficient search. The results obtained seem to agree with the increasing degrees of difficulty reported by psychophysical data from 23, 25 and 26, whose experiments were simulated above. Our experiments seem to agree also with the proposal that search is more efficient when objects are more dissimilar[6] and the continuum efficient-inefficient search found in the literature.[5]

3. Motion Model

Here we present a short description of the Motion Model and explain the main concepts and output conventions in order to be able to explain the experimental results. Mathematical details are omitted since they have been published elsewhere.[12]

3.1. *Description*

The Motion Model is a computational model of attention that works in the motion domain. As input it accepts a video stream in the form of sequences of images and is able to detect, localize and classify moving objects in the scene. The processing of information is inspired by biological research and therefore the computational structure of the model mimics some known properties of the monkey visual pathway. There are four distinct areas of the cortex that are simulated in the model: V1, MT, MST and 7a (Fig. 9). All these areas are known to participate in processing of visual information and specifically that which is perceived as motion. The model consists of 694 feature maps each of which encodes the whole visual field in a unique way. Those feature maps are organized into the areas based on their properties and areas are positioned in the form of a pyramid with information flowing from the input to the top of the pyramid and from the top back to the bottom providing feedback.

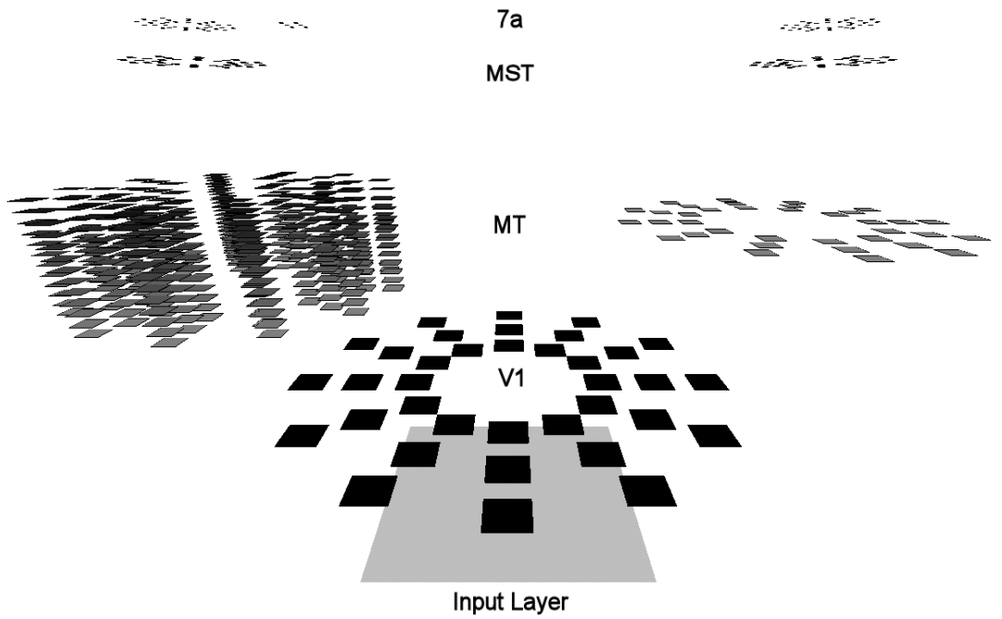


Fig. 9. Full hierarchy of the Motion Model. This shows the set of neural selectivities that comprise the entire pyramidal hierarchy covering visual areas V1, MT, MST, and 7a. Each rectangle represents a single type of selectivity applied over the full image at that level of the pyramid. The three rectangles at each direction represent the three speed selectivity ranges in the model. Position of a square around the circle represents direction selectivity. In area V1 the neurons are selective to 12 different directions and 3 different speeds (low, medium and high). Each area following area V1 has two parts. One where neurons are tuned to direction and speed, much like in V1 (the translational pyramid on the right) and the second part where neurons have more complex characteristics and are able to encode complex motion patterns, such as rotation, expansion and contraction (the spiral pyramid on the left). Colored rectangles in area MT represent particular angles between motion and speed gradient. MST units respond to complex patterns of motion. The 7a layers represent translational motion, complex motion, both as in area MST, plus radial and rotation without direction in the topmost set of six rectangles.

The internal architecture of the model is rather complicated and full description of it is beyond the scope of this paper (see Ref. 12). Here we present a brief description. From the input images, information channeled to area V1 which contains 36 feature maps. Each of those feature maps contains topographically arranged neurons which are tuned to one of the twelve directions and one of the three speeds. Spatiotemporal filters are used to model the selectivity of V1 neurons for speed and direction of local motion. The feature maps are positioned in three rings with twelve maps in each ring. The position in the ring corresponds to the preferred speed of the neurons (the outside ring represents high speed, middle ring — medium speed and the inside ring — low speed) (Fig. 9).

Area V1 projects onto area MT. Starting from area MT and on the processing is split into two pyramids: one that processes translational motion (the translational pyramid, Fig. 9 on the right) and the other that processes complex motion such as expansion, contraction and rotation (the spiral pyramid, Fig. 9 on the left). Area MT (Fig. 10) contains 36 feature maps of the translational pyramid and 432 feature maps in the spiral pyramid. The translational part of MT is in effect a blurred version V1 with smaller size and larger receptive fields. The complex motion part of MT, however, is a lot different from V1.

The aspect that is important to the current discussion is how the model processes complex motion patterns. Every point in the complex motion pattern moves with a unique velocity (i.e. the direction or the magnitude or both are different for every point). So as complex motion is processed by the model many different feature maps are activated by that motion. For example, the neurons of the area V1 encode only simple linear motion in 12 different

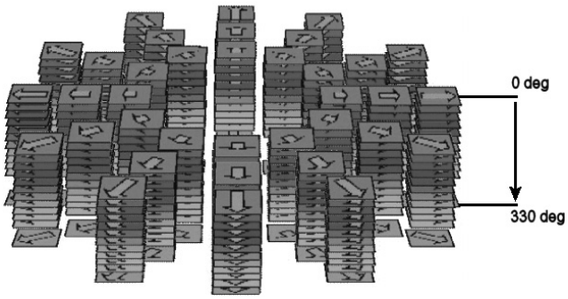


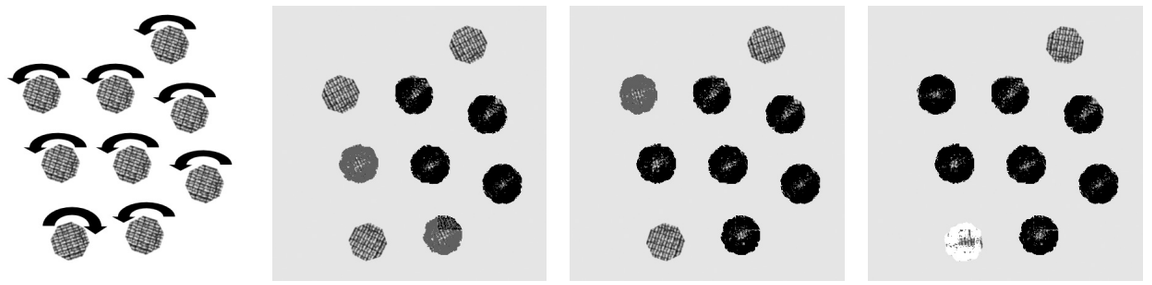
Fig. 10. The gradient part of MT is the largest area in the entire hierarchy. It consists of 432 feature maps and encodes every possible combination of 12 directions of motion, 3 speeds and 12 angles between direction of motion and the direction of the spatial derivative of local velocities. Here different gray values represent different different angles between local direction of motion and the spatial derivative of local motion (velocity gradient).

directions. Therefore all of V1 will have some activation since there are points moving in each of 12 directions encoded by V1. Further, in MT the moving object is decomposed into regions of common spatial derivatives of local velocity. The full representation of the complex motion is thus the conjunction of different features in the spiral part of area MT which encode different directions of motion but with the same angle to the speed gradient. Therefore the search for the target that exhibits complex motion among the complex motion distractors can be viewed as a conjunction search and can be expected to produce serial-like performance.

Area MST is located above MT. The translational part of MST contains 36 feature maps which receive their inputs from corresponding feature maps in translational part of MT. The size of feature maps of translational part of MST is smaller than that of MT and the receptive fields of the neurons are larger. The translational MST is essentially a blurred version of translational MT. The spiral part of MST receives its inputs from spiral MT maps it is able to combine together similar gradient patterns in order to determine if object in the scene exhibits rotation, expansion, contraction or combination of those motions. The spiral part of MST has 36 feature maps also organized in three rings. Belonging to a ring determines the speed of motion and the position around the ring identifies the type of motion.

At the very top we have area 7a. Both translational and complex motion part of 7a have the same conventions as those in MST but the size of each feature map is smaller. 7a has 6 more feature maps, three of those maps are tuned to rotational motion and the other three to radial motion.

This hierarchy is used to process the image sequences. As input enters the system the activations of neurons are computed and propagated through the pyramid to area 7a at the top. This results in bottom-up data driven activation of 7a neurons. The most salient location and feature map of the area 7a is selected using one of the rules outlined in (Ref. 12). This provides us with the knowledge of the type of motion that takes place in the scene (identified by the selected feature map) and also with a very course spatial location of the motion. In order to localize the motion pattern in the input image the neurons within the receptive field of the winning neuron of area 7a participate in modified Winner-Take-All (WTA) algorithm (Ref. 12) in order to obtain localization of



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

Fig. 11. Typical output of the search task. (a) Input example: target moving clockwise, distractors counter-clockwise. (a, b, c) The most conspicuous locations are attended first, the target is not found and the distractors are inhibited to allow for the new location to be examined. (d) the search is terminated when the target is found.

the signal in area MST. Each winner in MST initializes WTA process within its receptive field and this process is repeated through the rest of the hierarchy until the motion pattern is localized in the input image. The localized region is marked according to the preset color coding scheme where each color corresponds to a different type of motion.

3.2. *Motion visual search*

To test the performance of the Motion Model we carried out two experiments. First we examined how the model performs a standard visual search task. Secondly we replicated one of the psychophysical experiments by Ref. 27 that consisted of different moving patterns at different locations (Figs. 12 and 13).

3.2.1. *Singleton visual search*

Method: In the experiment we examined how the Motion Model performs a standard visual search task. We used a singleton design where each trial contained only one target and number of distractors was varied from trial to trial. Images of size 445×445 pixels contained one target and from 1 to 8 distractors. A typical input is shown on Fig. 11a (the arrows depict the direction of rotation and were not present in the input images). The target and distractor objects were identical textured octagons of 65 pixels in diameter. The target was rotating counterclockwise and distractors were rotating clockwise both with the angular speed of 3 deg/frame. The target and the distractors were randomly positioned on the white background without overlapping. Figure 11b, c and d show the progress of the search. Instead of measuring the reaction time for finding the target we counted the number of frames processed by the model until the target was localized. The least mean squares method was used to fit the straight line into the set of points.

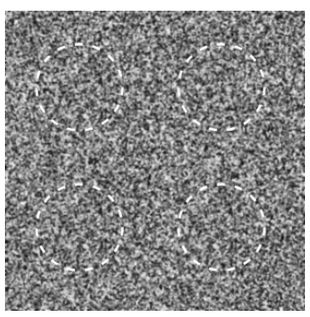
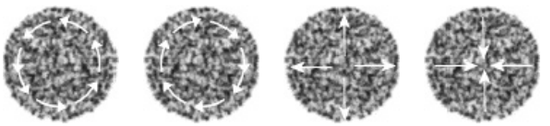


Fig. 12. Possible positions of the motion patterns in the input images.



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

Fig. 13. Motion patterns used in the experiment. (a) counterclockwise rotation, (b) clockwise rotation, (c) expansion and (d) contraction.

Results: Figure 14a shows how the time of detection of the target relates to the number of distractors present in the scene. The position of the points on the graph suggests a linear dependence. The straight line fit of the points has a slope of 1.34 frames/item and intersects with y-axis at 12.3 frames.

Discussion: We have shown in this chapter how Motion Model performs a standard visual search task. The equivalent of reaction time (RT) is expressed in the number of frames needed to find the target. The values appear to be linearly increasing as we increase set size, which seems to be in agreement with psychophysical data from Refs. 27 and 28. The typical output of the model is shown on the Fig. 9. We can see that objects are selected in groups rather than one at the time. This behavior is caused by the fact that the model is attending to the specific motion type at the specific spatial location. The location is defined by the receptive field of the winner neuron at the top of the pyramid. Therefore, every object or part of an object that lies within the attended receptive field and exhibits the attended motion will be selected and processed in parallel. Several other researchers proposed that multiple items can be processed in a single attentional fixation, see review Ref. 5.

3.2.2. *Thornton and gilden experiment*

Method: In experiment we compared the performance of the model with the human data by reproducing the experiment described in Ref. 27. The stimulus images consisted of a random noise background where every pixel was randomly assigned a value between 0 and 255 on which a motion patterns were superimposed. The motion patterns were

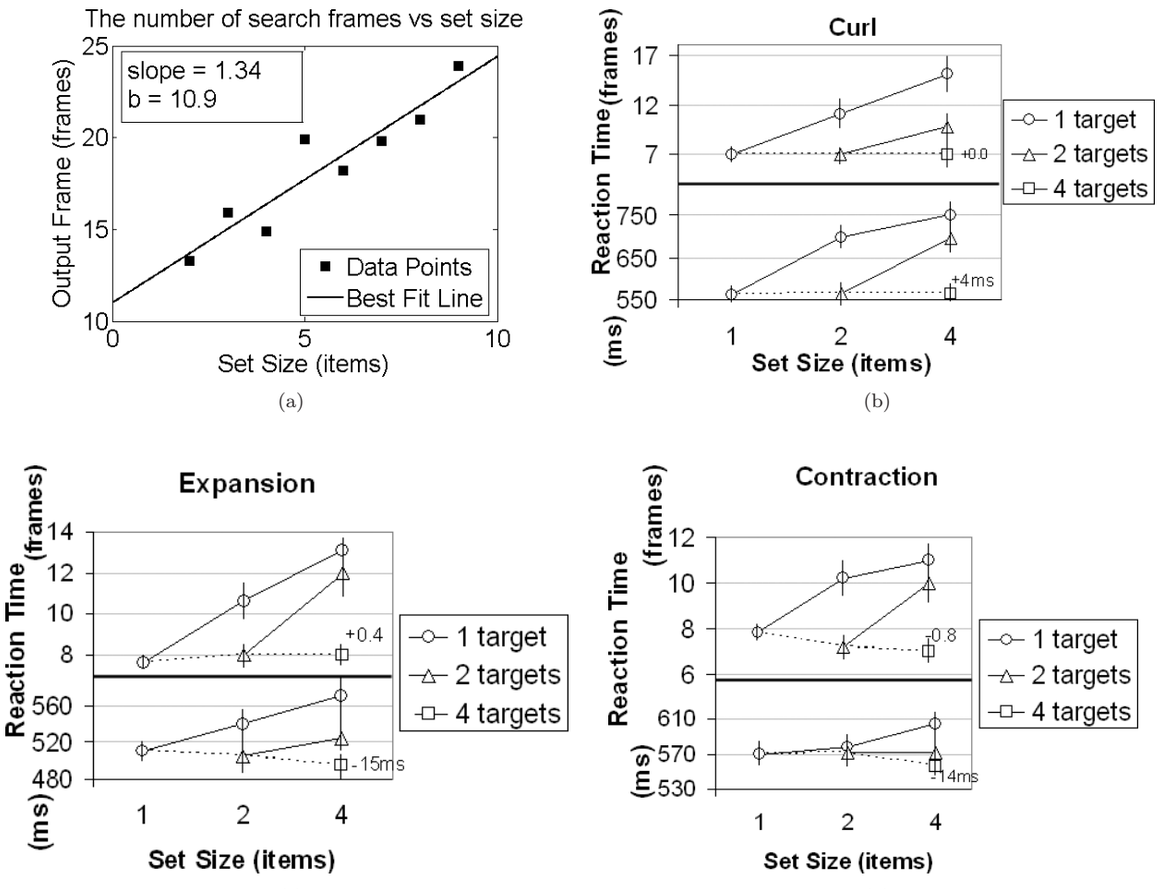


Fig. 14. Search Results (a) standard visual search for the stimulus in Fig. 9, (b, c, d) the model's performance on the stimuli used in Ref. 27. The top half of each graph shows the output of the model and the bottom half of the graph.

also comprised of dots of random intensity. Each dot was moved from frame to frame according to the motion transformation and a circular aperture was imposed on the motion patterns by repositioning the dots which rolled off the circle back into the aperture. The radius of the apertures was 43 pixels and the size of the image was 300 by 300 pixels. There were four positions where motion patterns could be placed, see Fig. 12.

For each type of motion there were six trial blocks with ten trails in each block. The number of targets and distractors was varied between blocks. The blocks contained either 1 target and 0 distractors, 1 target and 1 distractor, 1 target and 3 distractors, 2 targets and 0 distractors, 2 targets and 2 distractors or 4 targets and 0 distractors. The only difference between targets and distractors was the direction of motion. So for clockwise rotating targets the distractors were rotating counterclockwise and for expanding targets distractors were contracting and so on. After the motion patterns were placed on the background the whole image was smoothed by Gaussian filter with $\sigma = 0.75$. The types of motion patterns used in the experiment are shown on Fig. 12. The reaction time is expressed in terms of the number of frames needed to find the target.

Results: Figure 14b, c and d show the results of the experiment. The three graphs depict the model's performance on the stimuli used in Ref. 27. The top half of each graph shows the output of the model and the bottom half of the graph is the data reported by Ref. 27. The complex motion patterns produce

nearly linear dependence on the set size. The rotating motion shows the steepest slope among the complex motions which is in agreement with the human data.

Discussion: The results of this experiment show a lot of similarities between the output of the model and human performance on the visual search task. Although, no direct quantitative comparison can be done we can see that qualitative similarity is definitely present. The complex motion patterns seem to be handled by the model in a manner comparable to the human visual system. In the case of contraction, expansion and curl there is no decline in the RT as the number of targets increases, and there is a nearly linear rise of response times as the number of distractors increases. The curl patterns have the largest slope compared to other complex motions which is also in agreement with psychophysical data. Overall the comparison is qualitatively correct, an encouraging sign for the biological plausibility of the model.

4. Conclusions

Here we have shown how the Selective Tuning model can account for the visual search observations of a significant set of psychophysical experiments have been presented. Two very different set of stimuli have been used to test the model, one corresponding to colored shape objects while the other dealt with motion patterns. In each case, both feature singleton and feature conjunction image items can be correctly handled.

The model can also differentiate the different types of visual search experiments that have appeared along the years, showing a different efficiency no only between feature and conjunction searches but also more difficult searches (inefficient visual search) as the one described in Ref. 26. The behavior of the model agrees with well established models of visual search,[5,6] accounting for a continuum efficient-inefficient search related to the similarity between target and distractors.

The work is in stark difference to other seemingly related research (such as Refs. 8 and 9). Here the performance comparison is not eye movement based as in Ref. 8. They model bottom-up saliency and cannot include top-down effects of general knowledge while at the same time use tracking data that is confounded by such knowledge. Reference 9 also model bottom-up recognition with no need for attention and thus have no natural mechanism for serial search through a collection of stimulus items in a display. The contribution in this paper of mechanisms that can provide an explanation for visual search performance has the promise of enhancing performance of recognition algorithms in complex scenes.

References

1. J. E. Hoffman, in *Attention* (University College London Press, London, UK, 1995).
2. L. Itti, G. Rees and J. K. Tsotsos, *Neurobiology of Attention* (Elsevier Science, 2005).
3. A. Treisman and G. Gelade, *Cognitive Psychol.* **12** (1980) 97–136.
4. C. Koch and S. Ullman, *Hum. Neurobiol.* **4** (1985) 219–217.
5. J. Wolfe, in *Attention* (University College London Press, London, UK, 1995).
6. J. Duncan and G. Humphreys, *Psychol. Rev.* **96** (1989) 433–458.
7. J. Wolfe, K. Cave and S. Franzel, *J. Exp. Psychol. Human.* **15** (1989) 419–433.
8. L. Itti, C. Koch and E. Niebur, *IEEE T. Pattern Anal* **20**(11) (1998) 1254–1259.
9. M. Riesenhuber and T. Poggio, *Nat. Neurosci.* **2** (1999) 1019–1025.
10. H. Helmholtz, *J. Opt. Soc. Am.* (1924).
11. J. K. Tsotsos, S. Culhane, W. Wai, Y. Lai, N. Davis and F. Nuflo, *Artif. Intell.* **78** (1995) 507–545.
12. J. K. Tsotsos, Y. Liu, J. Martinez-Trujillo, M. Pomplum, E. Simine and K. Zhou, *Comput. Vis. Image Und.* **100**(1–2) (2005) 3–40.
13. A. Pasupathy and C. Connor, *J. Neurophysiol.* **86** (2001) 2505–2519.
14. S. Marcelja, *J. Opt. Soc. Am.* **70** (1980) 1297–1300.
15. J. Canny, *IEEE T. Pattern Anal.* **8**(6) (1986) 679–698.
16. D. Felleman and D. V. Essen, *Cereb. Cortex* **1**(1) (1991) 1–47.
17. E. Rolls and G. Deco, *Computational Neuroscience of Vision* (Oxford, New York, 2002).
18. L. G. Williams, *Percept. Psyhophys.* **1** (1966) 315–318.
19. J. D. Gould and A. Dill, *Percept. Psyhophys.* **6** (1969) 311–320.
20. P. Viviani and R. G. Swensson, *J. Exp. Psychol. Human.* **8** (1982) 113–126.
21. A. Allport, in *Foundations of Cognitive Science* (Ed. Posner, MIT Press Bradford Books, 1989).
22. D. Sharvit, J. Chan, H. Tek and BB. Kimia, *IEEE Workshop on Content-Based Access of Image and Video Libraries* (1998) 56–62. *http://www.lems.brown.edu/vision/software/index.html.*
23. A. Nagy and R. Sanchez, *J. Opt. Soc. Am. A* **7** (1990) 1209–1217.

24. K. G. Thompson, N. P. Bichot and T. R. Sato, *J Neurophysiol.* **93** (2005) 337–351.
25. N. Bichot and J. Schall, *Visual Nerosci.* **16** (1999) 91–99.
26. H. Egeth and D. Dagenbach, *J. Exp. Psychol. Human.* **17** (1991) 551–560.
27. T. Thornton and D. Gilden, *Cognitive Psychol.* **43** (2001) 23–52.
28. A. Hillstrom and S. Mantis, *Percept. Psyhophys.* **55**(4) (1994) 344–411.