Jaime A. Martins, J. M. F. Rodrigues & J. M. H. du Buf

Local Object Gist: Meaningful Shapes and Spatial Layout at a Very Early Stage of Visual Processing

1. Introduction

In his introduction, Pinna (2010) quoted one of Wertheimer's observations: "I stand at the window and see a house, trees, sky. Theoretically I might say there were 327 brightnesses and nuances of color. Do I have '327'? No. I have sky, house, and trees." This seems quite remarkable, for Max Wertheimer, together with Kurt Koffka and Wolfgang Koehler, was a pioneer of Gestalt Theory: perceptual organisation was tackled considering grouping rules of line and edge elements in relation to figure-ground segregation, i.e., a meaningful object (the figure) as perceived against a complex background (the ground).

At the lowest level – line and edge elements – Wertheimer (1923) himself formulated grouping principles on the basis of proximity, good continuation, convexity, symmetry and, often forgotten, past experience of the observer. Rubin (1921) formulated rules for figure-ground segregation using surroundedness, size and orientation, but also convexity and symmetry. Almost a century of research into Gestalt later, Pinna and Reeves (2006) introduced the notion of figurality, meant to represent the integrated set of properties of visual objects, from the principles of grouping and figure-ground to the colour and volume of objects with shading. Pinna, in 2010, went one important step further and studied perceptual meaning, i.e., the interpretation of complex figures on the basis of past experience of the observer. Re-establishing a link to Wertheimer's rule about past experience, he formulated five propositions, three definitions and seven properties on the basis of observations made on graphically manipulated patterns. For example, he introduced the illusion of meaning by comics-like elements suggesting wind, therefore inducing a learned interpretation. His last figure shows a regular array of squares but with irregular positions on the right side. This pile of (ir)regular squares can be interpreted as the result of an earthquake which destroyed part of an apartment block. This is much more intuitive, direct and economic than describing the complexity of the array of squares. Indeed, Pinna noted that such a formulation of what might have happened to the array of

GESTALT THEORY
© 2012 (ISSN 0170-057 X)
Vol. 34, No.3/4, 349-380

squares, a "happening," relates to "affordance" as introduced by James J. Gibson, a pioneer of ecological psychology also known as environmental psychology. Our brain has evolved, during hundreds of thousands of years, for perceiving and interpreting the natural world and, during only a few thousands of years, increasingly also a man-made world. It should therefore be no surprise, as Pinna (2010) found, that "everything has a meaning," because our brain has extensively learned to extract meaning. And we are thinking in terms of semantics, not in terms of low-level syntax, although it is now clear that part of our brain, at least the primary visual cortex, is devoted to low-level syntax. The link between low-level syntax and high-level semantics remains subject to research.

Pinna (2010) built on Gestalt and went one step further in the direction of less abstract but meaningful objects. We, here, take the opposite direction. By asking what is necessary to extract meaningful objects, it should be possible to go down to the level of Gestalt Theory. However, this only concerns principles such as good continuation and not past experience nor learned interpretations, because we also address abstract and therefore man-made objects, but only those without ambiguity and possibly at a very early stage in the visual system. For example, for detecting a square traffic sign – which is an explicit application of our research – one needs four edges connected at four corners, whereas a triangular one requires three edges and three corners. Occlusion of one of the square sign's corners does not necessarily lead to an interpretation of a triangle, nor does occlusion of one of the triangle's corners lead to a square. At a very early stage in visual processing, much below conscious reasoning on the basis of past experience, our neural circuits have also learned: to use and combine all available information in order to obtain the most robust and reliable solution. Although we may not be aware of this, it is crucial while driving a car at 120 km/h. In this sense, our research represents, as for Pinna's, a further step in Gestalt, but, unlike Pinna's, not necessarily at the level of conscious report. Although there is no doubt that geometric shapes like squares, triangles and circles are important at high semantic level, we will argue that they may also be important at a much lower level, for example in the dorsal "where" data stream for directing attention and eye/head control. In order to be able to understand why this may be the case, we need to explain some concepts of the visual system in more detail.

1.1 Attention, Spatial Layout and Local Gist

Our visual system is a very hierarchical one. The brain is divided into front and back parts, roughly at the central sulcus, with the front part "looking at" the back part (Crick & Koch, 2003). In the back part there is fast and massively parallel processing, from low-level syntax to invariant object representations, whereas the front part works at a slower pace, with serial processing involving covert and overt attention at semantic level. For example, experiments with different types

of "snake" patterns revealed that some may "pop out" instantaneously, which is an indication for parallel processing, whereas others require effortful scrutiny and serial processing (Houtkamp & Roelfsema, 2010). In this article we focus on the transition between low-level syntax and low-level semantics, using very elementary information such as colour. The goal is to develop a system for local gist vision: which types of objects are about where in a scene. This is necessary to bootstrap and guide, even alleviate, the processing in the ventral and dorsal data streams. These streams are known to serve two goals: the dorsal stream, also called the "where" or vision-for-action stream, is devoted to optical flow and stereo disparity, whereas the ventral stream, also called the "what" or vision-for-perception stream, is devoted to object recognition (Konen & Kastner, 2008; Farivar, 2009).

One problem is that precise object recognition in the ventral stream requires object segregation, but object segregation is only possible if the system already knows what the object is, assuming of course that objects are seen against complex backgrounds. Another problem is that object recognition is a sequential process: while fixating one object, its features must be routed to normalised object templates held in memory. This routing blocks the system until recognition has been achieved, after which the system is released for dealing with another object. Therefore Rensink (2000) proposed a non-attentional "scene schema" consisting of concurrent spatial-layout and gist subsystems which both drive attentional object recognition, all employing "proto-objects" resulting from low-level vision. However, gist vision addressed so far concerns global gist of entire scenes (Bar, 2004; Siagian & Itti, 2007; Ross & Oliva, 2010; Rodrigues & du Buf, 2011).

Global scene gist can be used to bias – select or exclude – object templates in memory in the matching process: when in a classroom it is not very likely that we see a horse. But global gist lacks localisation. On the other hand, when seeing a horse it is not very likely that we are in a classroom. Local object gist has the advantage of solving, or at least contributing to, Rensink's spatial-layout subsystem. Although both global and local gist can determine context, probably with a straight relation between them, local gist can solve important problems like a first and fast object categorisation, localisation and segregation, the latter being related to figure-ground organisation (Craft et al., 2007).

In what follows one should keep in mind that global gist vision is very fast: our brain is able to pre-categorise scenes in as few as 19-67 ms, but final scene recognition takes 100-200 ms, whereas object recognition takes 200-300 ms (Oliva & Torralba, 2006; Bar et al., 2006; Greene & Oliva, 2009). Local gist vision, also assumed to be very fast, is not meant for precise object recognition with conscious report; it may only be one preprocessing stage for guiding attention to meaningful locations while performing a specific task. For example, man-made objects with a simple shape repertoire include traffic signs, which is

one application that will be tested. The basic shape of a sign – circular, triangular or square – implies a certain function. Hence, the visual system of a car driver can already be alerted and biased when still being far away from a sign. After this, an attention window can be created and updated for covert or overt attention with eye fixations driven by the dorsal stream for determining, in the ventral stream, which sign it actually is when approaching it. This general organisation of the processing involved may appear intuitive, but in practice there are some complications, like the roles of episodic and procedural memories, i.e., learned observation and driving behaviour on often-driven roads, and, as we will see below, problems when dealing with multiple traffic signs which have been mounted together.

We know that the dorsal data stream for stereomotion (Peng & Shi, 2010) and motor control is faster than the ventral stream. Average activation latencies in the case of geometric shapes – which we address here – are 62 ms in the dorsal area LIP of the posterior parietal cortex and 101 ms in the ventral area AIT of the inferior temporal cortex (Lehky & Sereno, 2007); these areas are described below. Two key questions therefore are whether analysis of geometric shapes can occur in the dorsal stream, which is *not* devoted to object recognition, and whether this can occur at a very early stage, for example by directly employing simple information from retinal ganglion cells instead of much more complex information in cortical areas V1 and V2, but in both cases in areas which are very far from the front part of the brain, as referred to above, where semantic representations are handled. We must keep in mind that extraction of geometric shapes can be seen as a first object categorisation, and the latter can also be achieved at a higher level with coarse-to-fine-scale processing by combining information in the dorsal and the ventral streams (Rodrigues & du Buf, 2009b). For answering the two questions we need to go into more detail.

1.2 The Ventral and Dorsal Pathways

For a very recent overview of early processing, from retinae to lateral geniculate nuclei (LGN) to early visual cortical areas, we refer to Troncoso et al. (2011). After area V2, the dorsal where stream continues to areas MT, MST and other intermediate areas up to the posterior parietal cortex. MT neurons are selective to direction of motion, speed and binocular disparity. MST neurons convey more global information about a scene's structure and spatial relationships (Smith et al., 2006), including egomotion (Wall & Smith, 2008). The latter authors suggested that MST has a central role in guiding heading in macaques. All these processes would benefit from early object segregation such that motion and disparity are integrated and attributed to meaningful items in visual space, especially when also motion prediction is applied for estimating where objects are expected next. Motion prediction is a form of adaptation which can explain

the motion aftereffect, for example our illusion that a railway station moves after the train in which we sit has stopped. This may occur in area MT (Kohn & Movshon, 2003). These are indications that attention is not only a static process directed by complexity in the visual field, for example colour conspicuity, but a dynamic one involving motion and motion prediction. If these are processed at an early level, they (i) can control processing at a very low level, even down to the LGN (Kastner et al., 2006), and (ii) can have a bottom-up connection to the prefrontal cortex with two top-down attentional components from the prefrontal cortex, i.e., from PF46d (d from dorsal) to the posterior parietal cortex (see above) and from PF46v (v from ventral) to the inferotemporal cortex (see below) (Deco & Rolls, 2005).

Concerning the posterior parietal cortex, the highest visual area in the dorsal stream, Creem-Regehr (2009) presented an overview of the functionalities of its different sub-areas for sensory-motor planning and online control of eye, head, arm and hand, which includes pointing, reaching, grasping and even correct handling of tools. These functionalities are closely related to cognitive skills like imagining, gesturing and pantomime. For example, correctly grasping tools by their handles requires interaction between cognitive and action systems, where objects have sensory-motor affordances which guide behaviour on the basis of object structure, relevant goals, and the tools' known functions. It was even proposed to extend the dichotomous what/where organisation to a trichotomous what/where/how one, involving semantic memory (which tool), procedural memory (how to handle it), and attention (where is it). In general, attention and action can be related to three types of behaviour: (1) reflexive in the case of sudden events like a moving object or an unexpected sound, e.g., a cat's arousal, (2) covert or automatic in the case of frequent or repetitive tasks, and (3) overt or consciously controlled when a task requires close scrutiny. If motor actions are controlled by the superior colliculus SC (with binaural source localisation via the inferior colliculus to the SC), these behaviours may be based on three pathways: (1) from retinae straight to the SC, (2) from intermediate areas MT/MST to the SC, and (3) from the highest area, the posterior parietal cortex, but with input from area PF46d, to the SC. As we will see below, this is still speculative, but early object-centred segregation, localisation and attention through local gist vision can be very useful for steering most processes, especially in case of man-made objects with simple shapes.

The ventral what stream continues after area V2 to area V4 and other intermediate areas up to the inferotemporal cortex. Many V4 neurons code colour, also orientation, width and lengths of bars, and curvilinear as well as linear gratings. The main purpose of this stream is object recognition. Also in this stream it would make sense to extract geometric shapes at an early stage, for a first object categorisation like man-made items, for example traffic signs which

are always circular, triangular or square, with perspective projections to elliptic and trapezoidal shapes, vs. natural items like persons, animals, plants and trees. Given the role of the ventral data stream, it can be no surprise that different objects are coded there, but at a higher level. For example, Kiani et al. (2007), who studied 674 neurons in monkey inferotemporal cortex, clustered the neurons' responses to about 1100 natural and artificial object images belonging to 23 intuitive categories. They found that different categorical structures are represented by different subsets (or populations) of the neurons. Animate and inanimate objects are represented by different subsets. In the case of animate objects, different subsets are devoted to bodies, hands and faces, the latter being divided – in other subsets – into human and monkey faces. Bodies of humans, birds and four-limb animals clustered together, whereas lower animals like fish and reptiles formed another cluster. Interestingly, in case of artificial objects like furniture, lamps, kitchen utensils and home appliances, these categories were not represented in monkey inferotemporal cortex, with the exception of cars, despite the fact that the animals were raised in human houses and later in zoos.

Kiani et al. (2007) also performed a similar cluster analysis on the basis of responses of (simulated) simple and complex cells. This analysis revealed no such clusterings, which means that categorical structures are formed after area V1, i.e., in higher areas which group V1 elementary features into meaningful items. This grouping can be based on Gestalt rules like good continuity, and can resemble the grouping into simple geometric shapes of many man-made objects as will be explained below.

In earlier work we developed a framework for invariant object categorisation and recognition, assuming multiscale representations in the two streams: keypoints in the dorsal stream and lines and edges in the ventral one. Starting at a coarse scale, keypoints are used to route lines and edges of an unnormalised input object to those of normalised object templates in memory. This yields a first, fast, but coarse categorisation, after which information at progressively finer scales is added to refine categorisation until final recognition has been achieved (Rodrigues & du Buf, 2009b). Line, edge and keypoint detection are based on models of simple, complex and end-stopped cells in V1 and V2. One questionable assumption was that keypoints have a dominant role in the dorsal stream, and lines and edges in the ventral stream. This assumption was based on the now abandoned idea of a strict dichotomous organisation: in the meantime there is substantial evidence that the two streams are communicating at many if not at all levels, and that some processing may be common to both (Konen & Kastner, 2008; Farivar, 2009). The latter is supported by the processing of geometric shapes in both areas LIP (dorsal) and AIT (ventral), although activation in LIP is faster (62 ms vs. 101 ms) and LIP neurons are less selective to different patterns than AIT neurons (Lehky & Sereno, 2007).

One of the two key questions has been answered: the same geometrical shapes are processed in both data streams and the dorsal one is faster. The latter makes sense if the dorsal stream also goes, via areas MT and MST, to the superior colliculus for eye and head control. But above we wrote that area MT is fed by area V2, where lines, edges and keypoints may be extracted and processed for both data streams. That those two streams exist is now generally accepted, also the fact that the border between them may be rather fuzzy, but how about yet other data streams? For answering this question we must take a closer look at the retina.

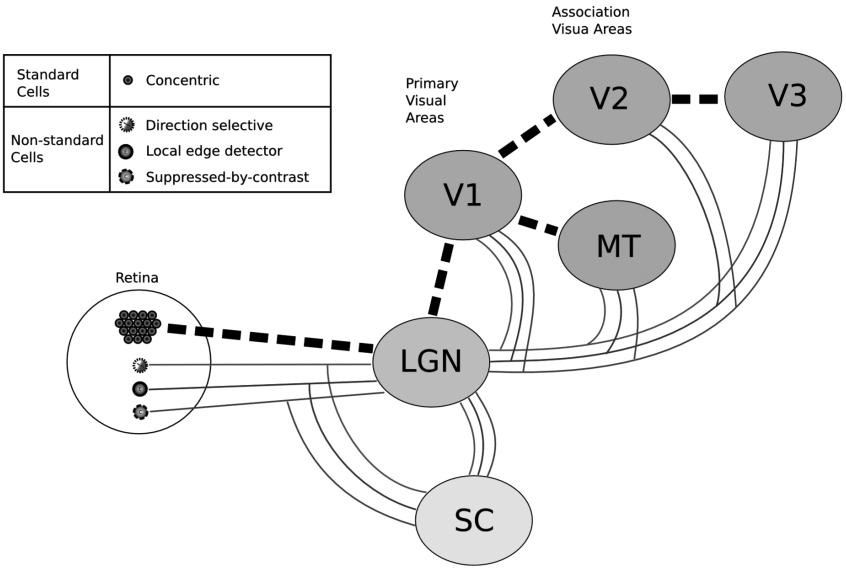

Fig. 1. Non-standard retinal ganglion cells and pathways. Figure adapted from Masland and Martin (2007).

1.3 Non-Standard Retinal Ganglion Cells

Cones (also rods) are photoreceptors which sample the image projected onto the retina. Horizontal and other cells combine the samples into concentric bandpass functions, the ON and OFF signals, often visualised as Mexican hat or Difference-of-Gaussian (DOG) functions. Retinal ganglion cells transmit these signals to the LGN in the thalamus, which relays them on to the primary cortex at input level V1 where simple, complex and hypercomplex cells reconstruct anisotropy for building multi-scale line, edge and keypoint representations. Most of the ganglion cells feature high spatial resolution for the ventral data stream, fewer cells are devoted to motion for the dorsal stream, and both types are called standard retinal ganglion cells. It is now known that there also are, although even less, non-standard ganglion cells devoted to other functions: these are direction-selective cells, local edge-detection cells and suppressed-by-contrast

cells. All these project not only on the superior colliculus for direct motor control, but also on the LGN and higher areas MT, V2 etc. (Masland & Martin, 2007); see Fig. 1. In other words, the retina is much more intelligent than thought before, although much work remains to be done in order to better understand the role of the non-standard cells. Such cells and their pathways may explain the phenomenon of blind sight: some persons or primates without visual cortex, who are effectively blind, can nevertheless avoid obstacles. A monkey named Helen was even able to detect, localise and discriminate visual objects (Stoerig & Cowey, 1997). Hence, there must exist more pathways which complement the dorsal and ventral ones. Another one is that for face detection and recognition, because faces are not normal 3D objects which can be arbitrarily rotated and they play an important role in social contacts. Their processing must be very fast and may circumvent the pathway for normal 3D objects (Biederman & Kalocsai, 199The second key question has now also been answered: if static and moving edges are already detected in the retina and this information is conveyed by non-standard retinal ganglion cells, either directly or indirectly, to areas like MT and the superior colliculus which are involved in motion prediction and eye-head control, it makes sense to assume that bottom-up attention focuses on entire (moving) objects. Such areas probably have sufficient computational resources to apply some principles of Gestalt Theory, like good continuation in case of edges which are partially occluded. Also for figure-ground, since motion and disparity may easily separate a moving object from its background and for the extraction of elementary geometric shapes like squares, triangles and circles which are abundant in the man-made world. Hence, local gist vision of meaningful objects may be obtained or at least prepared at a very early stage, and even in the dorsal data stream.

1.4 Behavioural Studies Involving Geometric Shapes

The geometric shapes used by Lehky & Sereno (2007) include a square, a triangle and a circle, which in most countries are used for traffic signs. One would expect that studies related to traffic safety underpin the importance and meaning of such shapes, including the speed of detection of and discrimination between them. Indeed, there exist studies, but mainly behavioural ones addressing eye movements and fixations. For example, Luoma (1992) found that, while driving a car, traffic signs are analysed with an average glance duration of 500 ms, and glances as short as 100 ms may be enough to identify sign shape and colour. The average fixation time of 500 ms was confirmed by Martens & Fox (2007), but: (a) total fixation times including repeated fixations of the same sign were 500 ms with a standard error also close to 500 ms, such that total time varied between 0 and 1 s; (b) the total number of fixations ranged from 0 to 2; (c) there was a large variability between signs with different shapes, between individually

mounted signs and signs mounted next to one or two other signs; (d) there was almost no difference between real driving and looking at a video; and (e) fixation times became shorter when the road became more familiar to the driver. The last effect may be due to the roles of procedural and episodic memories while simultaneously observing the road and controlling the car. Interestingly (or alarmingly?), shortest fixation times were measured for round and triangular signs with a red border (pedestrian; speed limit) when these were mounted next to two other signs, and for an individual white round sign (end speed limit), all also with the smallest number of fixations. Fixations were measured at distances smaller than 250 m, with a normal speed of perhaps 50 km/h which means a maximum visibility of 18 s. This may imply (1) that most drivers were overloaded and simply ignored multiple signs, paying more attention to individual signs, or (2) for most drivers one brief glance at some distance was enough to grasp shape and meaning. If the ventral data stream is required to separate and analyse combined signs with one or two fixations, this stream may be too slow (option 1), whereas the dorsal stream may analyse fast individual signs (option 2).

Unfortunately, Martens & Fox (2007) only used signs with a symbol in the centre, which is a complication when one is only interested in the effect of a sign's shape. Karttunen & Häkkinen (1982) studied the discrimination of traffic signs in peripheral vision. They used 10 signs, only two of which without a symbol in the centre: a triangle and a circle. In all experiments these scored highest, i.e., completely correct discrimination of shape, colour and symbol, the latter lacking of course. The size of the signs was 4 degrees, which in practice corresponds to an observation distance of about 10 m, and the presentation time was 125 ms. Hence, the size was comfortable, but the presentation time was too short to identify also the symbol, if present. These results were obtained at peripheral angles beyond 30 degrees where retinal resolution is reduced; at angles below 30 degrees the differences between the scores were less. If flanked by two other signs (one above and one below), the same two signs also scored highest, but a bit lower if compared to the non-flanked condition. The discrimination results of Karttunen & Häkkinen (1982) seem to confirm the eye-fixation results of Martens & Fox (2007), namely that symbols in the centres of traffic signs complicate recognition. This effect might be due to the fact that fast and low-level gist of geometric shapes in the dorsal pathway is a possibility, but it has limitations because it is not intended for the recognition of more complex patterns. The latter require (para-)foveal vision and probably processing in the ventral pathway. This is supported by Lehky & Sereno (2007), who found that activation in the dorsal area LIP is faster than that in the ventral area AIT, and that LIP neurons are less selective to different patterns than AIT neurons.

1.5 From Gestalt Theory to Application

Gestalt Theory has been and still is a useful paradigm to understand how and why we perceive certain structures, from very simple and elementary patterns to rather complex ones which reveal meaning on the basis of learned interpretations (Pinna, 2010). Since our visual system is still much more efficient and reliable than most systems developed in computer vision, for example in robotics, it makes sense to apply our knowledge of the visual system, via development of advanced models, to real-world problems. Autonomous service robots, for example, must be able to deal with cluttered scenes, often containing complex objects which must be recognised and manipulated. Apart from scene and object complexity, digital images captured by modern cameras may have a good resolution, but they still contain noise due to digitisation and environmental factors like edges caused by nonuniform illumination. We will mention only two approaches. The first one is a nice example of the explicit application of the grouping rules of Gestalt Theory. The second does not apply such rules, but it represents an ideal case which could profit from advanced models of local gist and attention.

In their recent paper, Pugeault et al. (2010), who already applied grouping rules for the detection of lines and edges in their previous work, combined line and edge detection in 2D images with stereo disparity. Using Gestalt's grouping rules as constraints, they showed that 2D detection combined with 3D information leads to much more robust detection, especially in image regions where many features are very close and provide ambiguous information caused by local complexity. In other words, the application of constraints based on grouping rules is able to disambiguate such local information, leading to more consistent and complete 3D object and scene representations. Given the facts that the cortical hypercolumns in area V1, originating from retinotopic projections of the left and right eyes, are very close there, and that simple and complex cells in V1 serve to code lines and edges, it is very likely that our visual system extracts 3D shape information already in V1 and attributes depth to (mainly vertical) lines and edges. In addition to exploiting optical flow, this facilitates segregation (figure-ground) into meaningful and 3D objects.

Faubel & Schöner (2009) developed a system for simultaneous localisation and recognition of objects in a scene. They apply a circular Gaussian "receptive field" (RF) to extract a colour histogram of pixels with saturated colours, which is invariant to 2D rotation, and an edge-orientation histogram, which is cyclically rotated after 2D rotation. These histograms are complemented by a shape descriptor resulting from maximum pooling. First, objects are learned by positioning them in the RF, where each object can be represented by multiple views and therefore multiple histograms.

Then, a scene is analysed by covering the entire image with partly overlapping

RFs. For the sake of simplicity we will explain the analysis here using only one histogram, for example colour. In the first step, the histograms of all RFs are multiplied by weight factors and summed, which yields one histogram. This histogram is correlated with the histograms of all objects in memory, and the correlation factors are subjected to competition in order to suppress unlikely objects. Then all histograms in memory are multiplied by the reduced correlation factors and summed, which yields again one histogram. This histogram is correlated with all histograms of the RFs in the image, and the correlation factors are subjected to competition, as before, but now to suppress unlikely object positions. The reduced correlation factors are used as weight factors in the first step. Hence, the analysis is done in a closed loop from input space to memory and back to input space. This entire process is controlled by a dynamic neural field system over space and time, such that the solution can converge to one object at one position.

The closed loop resembles the processing in the visual system, with bottom-up and top-down projections which converge to a stable solution on the basis of adaptation or plasticity at different levels. First localisation is obtained (where) and then precise object recognition with pose (view) estimation (what), in which different weight factors of the histograms and shape descriptors are applied. The main problems, of course, are that only one object can be dealt with at any time, that objects may be partially occluded, that the sums of the histograms of two different objects may resemble the histograms of one other object, and that a complex background may lead to false positives. However, our visual system must deal with the same problems. As explained before, our visual system serially fixates regions on the basis of saliency, conspicuity and attention, the latter also driven by gist vision. Therefore, the system as developed by Faubel & Schöner (2009) could be modified such that it first applies object categories to an entire scene – which types of objects are about where – and then localised attention windows for identifying the objects in those windows instead of in the entire scene, but, as for the visual system, this step could be done serially.

1.6 Summary and Outlook

In summary, there is evidence for differences but also similarities of the processing in different pathways in our visual system. The times of 62 and 101 ms in areas LIP and AIT as measured by Lehky & Sereno (2007) are *activation* times, i.e., onsets of neural activities. After the onsets, activities of neurons in both areas reach a peak and then decay, but they remain active until about 500 ms after stimulus onset. This indicates that the sets of measured neurons are part of bigger populations which serve to process the input patterns, but it is not yet quite clear what these populations do and how they do it, especially in the posterior parietal cortex in the dorsal stream. Different populations in the

inferior temporal cortex in the ventral stream code categorical structures like bodies and hands, in principle the basic information which must be combined to detect and identify entire objects. In any case, it seems that local gist vision, at least involving elementary geometric shapes of man-made objects, is possible in early vision and also in the dorsal data stream. Specifically, below we focus on man-made objects which are dominated by a simple shape repertoire: squares, rectangles, trapeziums, parallelograms, triangles, circles and ellipses. It is shown that such shapes can be detected by a hierarchy of a few cell layers, with strictly bottom-up or data-driven processing. As we will see, straight bars and curves known from Gestalt Theory must be complemented by corner information for such shapes, and all information must be combined by a few grouping rules which specify each shape.

The rest of this paper is organised as follows: the next section deals with different cell layers to obtain low- and mid-level geometry. Section 3 explains shape retrieval using mid-level geometry. In Section 4 we discuss our approach and lines for future research.

2. Low and Mid-Level Geometry

In this section we explain the process of preparing shape retrieval by low- and mid-level geometry. This is a two-fold process: we first construct a hierarchical *cell layer map* which encodes local geometric primitives by grouping cells: the primitives' *type* and *orientation*. Then, this information is used to detect geometric shapes based on spatial relationships between the grouping cells.

2.1 Cell-Layer Map Construction

We postulate a bottom-up hierarchy of cell layers in which each layer serves a specific purpose: (1) colour normalisation and boundary enhancement which mimic double-opponent cells (Bomberger & Schwartz, 2005), (2) detection of salient image points and regions, (3) enhancement of the most salient features, (4) determination of feature properties like orientation, aperture and curvature, (5) feature type assignment, (6) corner grouping cell condensing, and (7) object shape identification. Below the layers are explained in detail.

2.1.1 Light source normalisation layer

The input image $I(x,y)$ is first colour corrected $(I_{cc} = f_{cc}(I))$, taking into account the geometry and temperature of the light sources. Let each pixel P_i of image $I(x,y)$ be defined as (R_i,G_i,B_i) and (L_i,a_i,b_i) in the RGB and Lab colour spaces, with $i = \{1...N\}$, N being the total number of pixels in the image. We use Lab as an informal abbreviation for the CIE 1976 (L^*,a^*,b^*) colour space.

We first process the input image I_{in} using the two transformations described by Finlayson et al. (1998) and Martins et al. (2009), as shown below in Eq. 1 and Eq. 2, both in RGB colour space. This method applies iteratively steps A and B $(P_i^A \rightarrow P_i^B \rightarrow P_i^A \rightarrow P_i^B \rightarrow ...)$, step A being local and step B being global, until colour convergence is achieved, usually after 4–5 iterations. Each individual pixel is first corrected in step A for illuminant geometry independency, i.e., chromaticity. If $S_i = R_i + G_i + B_i$ then

$$P_i^A = (R_i / S_i, G_i / S_i, B_i / S_i). \quad (1)$$

This is followed in step B by global illuminant colour independency, i.e., grey-world normalisation. If $S_R = 1/N \cdot \sum_{j=1}^{N} R_j$, and similarly S_G and S_B, then

$$P_i^B = (R_i / S_R, G_i / S_G, B_i / S_B). \quad (2)$$

After the process is completed, the resulting RGB image is converted to Lab colour space and the a_{cc} and b_{cc} components, where subscript *cc* stands for colour-corrected, are combined in I_{cc} together with the *unmodified* L_{in} channel from the input image I_{in}, as depicted in Fig. 2. The main idea for using the Lab space is that it mimics double-opponent colour cells found in human vision, making it more useful for determining the conspicuity of borders between regions. The reason for using the L_{in} component instead of the L_{cc} one is that, as observed by Finlayson et al. (1998), the simple and fast repetition of steps A and B does a remarkably good job. In fact, it does the job too well because all gray pixels (with values R=G=B from 0 to 255) end up having R=G=B=127. In other words, all information in gray image regions will be lost. Hence, we use only the colour corrections of the method, while keeping the lightness channel of the original image for maintaining all the image's details. Figure 3 shows three results of colour correction applied to the traffic sign image, from top-left to top-right: original image, modified image with a blue tint (R -12%, G +4% and B +50%), and modified image with a warm white balance. The three results are shown below the input images. As can be seen, colour correction yields very similar images despite the rather large differences in the input images. Colour correction as explained above simulates colour constancy as employed in our visual system (Hubel, 1995). Summarising, the initial I_{in} image in RGB is normalised to I_{cc} and then converted to the colour space $L_{in}a_{cc}b_{cc}$.

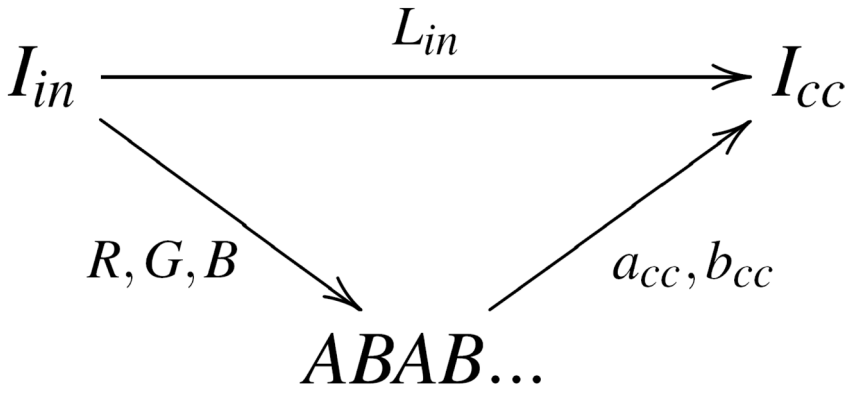

Fig. 2. Colour normalisation: the output image I_{cc} consists of the I_{in} component from the input image I_{in} and the a_{cc} and b_{cc} components after applying the algorithm of mFinlayson et al. (1998).

2.1.2 Adaptive colour filtering layer

After colour correction, image regions are smoothed for removing redundant information which is not necessary for shape detection, while preserving the regions' boundaries. The smoothing is done using an adaptive filter $\Gamma(x, y)$, with separable and equal $\Gamma^H = \Gamma(x)$ and $\Gamma^v = \Gamma(y)$ components for horizontal and vertical filtering. Each component consists of a centred DOG

$$F_{1,2}(x) = N_1 \cdot \left\{\exp\left(-x^2/2\sigma_1^2\right) - \exp\left(-x^2/2\sigma_2^2\right)\right\}, \tag{3}$$

which is split into $F_1(x<0)$ and $F_2(x>0)$, and a centred Gaussian which is *not* split,

$$F_3(x) = N_2 \cdot \exp\left(-x^2/2\sigma_2^2\right), \tag{4}$$

taking $\sigma_1 >> \sigma_2$. N_1 and N_2 are normalisation constants which make the integrals of all three functions equal to one. The three functions implement a group of three summation cells at the same position, but with different dendritic fields in the colour-opponent channels *a* and *b* of Lab colour space. F_3 yields the excitatory response of a cell with an *on-centre* dendritic field, whereas $F_{1,2}$ yield the excitatory responses of two cells with *off-centre* dendritic fields. From the three cell responses $R_{1,2,3}$ we first compute the contrast *C* between the left (R_1) and right (R_2) responses,

$$C = \left|(R_1 - R_2)/(R_1 + R_2)\right|. \tag{5}$$

Then, using the contrast *C* and the minimum difference between the on-centre response R_3 and the left and right off-centre ones R_1 and R_2, the response R of an output cell is determined by

$$R = \begin{cases} CR_1 + (1-C)R_3 & \text{if } |R_1 - R_3| < |R_2 - R_3| \\ CR_2 + (1-C)R_3 & \text{otherwise.} \end{cases} \tag{6}$$

In words, if the local contrast is low, as in almost homogeneous regions, the filter support is big, but if the contrast is high, at the boundaries between regions, the filter support is small. The adaptive filtering is applied to I_{cc} at each pixel position (x,y), first horizontally with Γ^H and then vertically with Γ^V,

$$I_{ci}(x,y) = \Gamma^V\left[\Gamma^H\left[I_{cc}(x,y)\right]\right], \tag{7}$$

where subscript ci stands for colour-improved. Results after both steps, for the illustration image, are shown in Fig. 3 (bottom-left). It can be seen that the filter not only preserves boundaries, but also sharpens blurred ones. In our experiments we obtained good results with $\sigma_1 = 7$ and $\sigma_2 = 3$, and adaptive filtering in horizontal and vertical directions was sufficient to sharpen blurred boundaries even with oblique orientations. Furthermore, the processing is very fast because the three filter functions need only be computed once.

Fig. 3. Top and middle: colour illuminant and geometry normalisation; input images (top) and respective results (middle). Bottom, left to right: adaptive colour-region filtering, border saliency by colour conspicuity, and non-maximum suppression.

2.1.3 Colour conspicuity layer

Following the idea of Martins et al. (2009), *conspicuity* $\Psi(x,y)$ is defined as the maximum difference between colours in I_{ci} at four pairs of symmetric positions at distance *l* from (*x*, *y*), i.e., on horizontal, vertical and two diagonal lines. However, here we apply a new concept of conspicuity directly to I_{ci}, effectively discarding the need for a previous edge-filtering step.

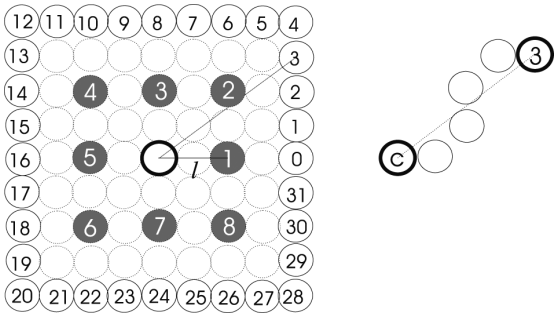

Fig. 4. Left: cell clusters, in gray the four clusters of gating cells at positions (1,5), (2,6), etc. used for colour conspicuity. The cells numbered from 0 to 31 serve orientation, curvature and connectivity detection. Right: an example of a possible path between the centre cell *c* and cell 3.

Figure 4 (left, in gray) shows the positions of the clusters of gating cells. If the gating cells are called G_i, opposite pairs are (G_i, G_{i+4}), with $i = \{1,...,4\}$, for example (G_1, G_5) and (G_4, G_8). We define conspicuity Ψ as the maximum Euclidean distance between the pairs of colour triplets (*L*, *a*, *b*) of the four pairs of opposite cells around (*x*, *y*),

$$\Psi_{lab}(x,y) = \max_{i=1}^{4} \sqrt{\sum \left(I_{ci}^{L,a,b}\left(\vec{x}_i\right)^2 - I_{ci}^{L,a,b}\left(\vec{x}_{i+4}\right)^2 \right)}. \tag{8}$$

The result of this layer as shown in Fig. 3 (bottom-middle) was obtained by using a distance $l = 1$ and a threshold at 0.4 of $\max(\Psi_{Lab})$ in the entire image.

2.1.4 Non-maximum suppression layer

In this cell layer Ω non-maximum suppression is applied in order to extract the positions where $\Psi(x,y)$ has a local maximum in horizontal, vertical and diagonal directions, in 3×3 neighbourhoods. As in the previous layer, this is achieved by four oriented cell clusters plus one grouping cell at the output. Mathematically,

$$\Omega(x,y) = \begin{cases} \text{ON} & \text{if } \left(\Psi(x_i,y_i) > \Psi(x_{i-1},y_{i-1}) \wedge \Psi(x_i,y_i) > \Psi(x_{i+1},y_{i+1})\right) \\ & \quad \vee \left(\Psi(x_i,y_i) > \Psi(x_{i-1},y_{i+1}) \wedge \Psi(x_i,y_i) > \Psi(x_{i+1},y_{i-1})\right) \\ & \quad \vee \left(\Psi(x_i,y_i) > \Psi(x_{i-1},y_i) \wedge \Psi(x_i,y_i) > \Psi(x_{i+1},y_i)\right) \\ & \quad \vee \left(\Psi(x_i,y_i) > \Psi(x_i,y_{i+1}) \wedge \Psi(x_i,y_i) > \Psi(x_i,y_{i-1})\right) \\ \text{OFF} & \text{otherwise.} \end{cases} \tag{9}$$

Results for the illustration image are shown in Fig. 3 (bottom-right).

2.1.5 Local feature layers

In order to extract meaningful information from layer Ω it is necessary to analyse local geometric relations between adjacent activated cells. We use three parallel cell layers Θ, Υ and Λ, which are dedicated to orientation, curvature and connectivity, respectively.

The **orientation layer** Θ encodes edge orientations in local neighbourhoods. Each active cell $(\Omega(x,y) = \mathrm{ON})$ triggers a cluster of 32 cells, each with two dendrites: one at the Ω cell's position (*x*, *y*) and one at a (discretised) distance of four cells (pixels) around that position, for a total of 32 orientations. This is illustrated in Fig. 4 (left), with orientations *n* numbered 0 to 31. Those of all 32 cells with dendritic input equal to 2 are excitated and the others are inhibited. Excitated cells provide the output of the Θ layer, the cells themselves implicitly coding all detected local orientations in the Ω layer. The dimension of the Θ layer is 32 times that of the Ω layer to accomodate all possible local edge orientations. Mathematically, the orientation equals

$$\phi_n = \begin{cases} \arctan(\theta/4) & \text{if } 0 \le n < 4 \\ \arctan(4/|8-\theta|) & \text{if } 4 \le n < 8 \\ \pi/2 & \text{if } n = 8 \end{cases} \tag{10}$$

in the case of the first quadrant $0 \le n \le 8$ and similarly for the other quadrants.

The **curvature layer** Υ is composed of clusters of curvature detection cells. These cells are also triggered by active output cells of the Ω layer and they also analyse active output cells of the Ω layer at a distance of about four cells (pixels). However, instead of combining the centre position (*x*, *y*) and one on a circle around it as in the Θ layer, they combine pairs of positions at near-opposite orientations on the circle (active Ω cells at exactly opposite orientations from the centre indicate zero curvature). In addition, since evidence for different local curvatures must be combined by grouping cells which determine the average curvature, evidence for curved edges on, for example, the left and right sides of (*x*, *y*) cannot be grouped because the average may be close to zero. Therefore, information on all semi-circles is grouped and the output of layer Υ is composed of 16 times 2 cells. Mathematically, the curvature model resembles computing the cluster curvature index $C_i(x,y)$ of all intersections of lines perpendicular to lines between all point pairs on a semi-circle, i.e., the mean Chebyshev distance between the *N* intersections (x'_k, y'_k) and the centre (*x*, *y*):

$$C_i(x,y) = \frac{1}{N}\sum_{k=1}^{N} \max\left(|x - x'_k|, |y - y'_k|\right). \tag{11}$$

The **connectivity layer** Λ also analyses the output of the Ω layer, but it employs the outputs of the Θ and Υ layers. Active or excitated output cells in those layers trigger clusters of grouping cells in the Λ layer: all detected orientations (Θ) and curvatures (Υ) trigger grouping cells which check whether there are active output cells in the Ω layer which connect the centre position (*x*, *y*) with the corresponding active cells on the circle around the centre. If so, output cells of the Λ layer are activated and these signal connectivity in the corresponding orientations.

In addition to the three layers described above, there is on top of Θ a $\hat{\Theta}$ layer which uses information of the connectivity layer Λ. This layer groups all detected orientations (active Θ cells) with confirmed connectivity (active Λ cells) for determining the *average orientation* on the entire circle. If there are opposite orientations, the average orientation can be orthogonal, but this information will be combined with other information at a higher level for distinguishing between corners and bars. The $\hat{\Theta}$ layer is complemented by a $\breve{\Theta}$ layer which determines the *angular aperture*, i.e., the spread of all orientations around the average orientation. If the aperture is small, this is evidence for a corner, but a large one indicates a continuous structure like a bar. Mathematically, the angular aperture $\breve{\phi}(x,y)$ and the average orientation $\hat{\phi}(x,y)$ are defined by

$$\breve{\phi}(x,y) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\phi_i - \phi_j\right| \wedge i \neq j \wedge 180^\circ \geq \left|\phi_i - \phi_j\right| > 22.5^\circ, \tag{12}$$

$$\hat{\phi}(x,y) = \frac{1}{N}\sum_{i=1}^{N}\phi_i, \tag{13}$$

with ϕ_i and ϕ_j the angles of active cells in the Θ layer. The value of 22.5° $(2 \cdot 360^\circ / 32)$ is an empirically chosen threshold angle for detecting corners.

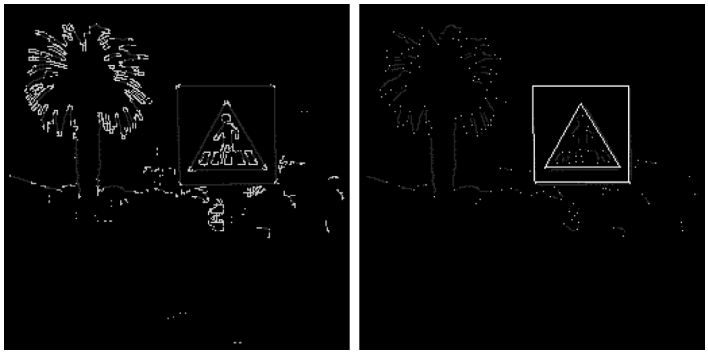

Fig. 5. Left: cell layer Ξ^{BCo} for the illustration image before corner condensing. Right: after corner condensing.

2.1.6 Mid-level geometry

This layer Ξ serves to translate the local low-level features into meaningful geometric primitives: straight bars, curves and corners. This is achieved by combining the information in the three previous cell layers Θ, Λ and Υ, more specifically, in $\hat{\Theta}$, $\breve{\Theta}$ and Υ. But the processing is still local: layer Ξ assigns a geometric primitive to each active Ω cell by a one-to-one mapping. Layers $\hat{\Theta}$ and $\breve{\Theta}$ provide orientation angles and apertures for corners and straight bars, whereas layer Υ provides curvature information. There are two parallel Ξ layers, Ξ^{BCo} and Ξ^{BCu}, the first for *bar-corner* cell clusters and the second for *bar-curve* cell clusters, with the purpose of fast shape processing in the further steps.

Mathematically, the angle aperture $\breve{\phi}(x,y)$ and curvature index $C_i(x,y)$ are used to determine which feature type $\{t_1,t_2\}$ will be assigned to cells $\xi^{t_1}(x,y)$ and $\xi^{t_2}(x,y)$,

$$\xi^{t_1}(x,y)=\begin{cases}\text{Bar} & \text{if } \breve{\phi}(x,y)\geq 120^{\circ}\\ \text{Corner} & \text{if } \breve{\phi}(x,y)<120^{\circ}\end{cases} \tag{14}$$

$$\xi^{t_2}(x,y)=\begin{cases}\text{Bar} & \text{if } \breve{\phi}(x,y)\geq 120^{\circ}\vee C_i(x,y)\leq 3\\ \text{Curve} & \text{if } \breve{\phi}(x,y)<120^{\circ}\wedge C_i(x,y)>3.\end{cases} \tag{15}$$

An example of the assignment (layers Ξ^{BCo} and Ξ^{BCu}) is shown in Fig. 5 (left), with white pixels being corner cells, gray pixels being curve cells and dark gray pixels being bar cells.

An overview of the processing is illustrated in Fig. 6, with the different cell layers in the case of corner detection. Non-maximum layer Ω shows the 9×9 neighbourhood with cell numbering as in Fig. 4. Conspicuity layer Ψ is a 9:1 mapping of layer I_{ci}. Similarly, non-maximum suppression layer Ω is a 9:1 mapping of layer Ψ. The corner in layer Ω is checked for curvature in Υ, for orientation in Θ, and for connectivity in Λ. It is then represented as a bar in layer Ξ^{BCu} and as a corner in layer Ξ^{BCo}.

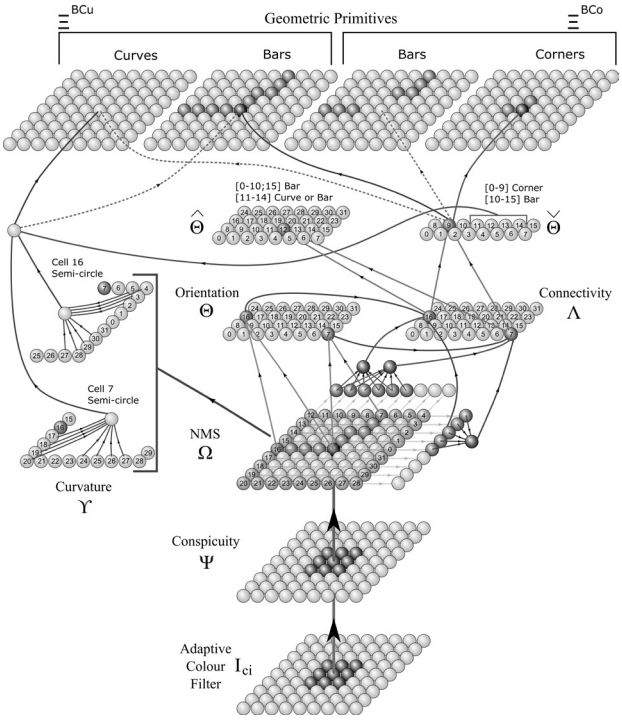

Fig. 6. Illustration of the whole process with the different cell layers in the case of corner detection. Conspicuity layer Ψ is a 9:1 mapping of layer I_{ci}. Non-maximum suppression layer Ω is a 9:1 mapping of layer Ψ. Non-maximum layer Ω shows the 9×9 neighbourhood with cell numbering as in Fig. 4. A corner in layer Ω is checked for curvature in Υ, for orientation in Θ, and for connectivity in Λ. It is represented as a bar in layer Ξ^{BCu} and as a corner in layer Ξ^{BCo}.

Corner cells are subjected to one more processing step, *condensing*, similar to Krüger et al. (2003). First, all active corner cells that have six or seven inactive neighbouring cells are inhibited. Second, all groups of cells $\xi(x,y)$ in layer Ξ^{BCo}, in a 7×7 neighbourhood, are condensed into a single "centre-of-gravity'' cell, with orientation and aperture angles equal to the averages of the group. This facilitates and speeds up further processing of corners for final shape recognition. An example, for the illustration image, is shown in Fig. 5 (right), where groups of corner cells have been replaced by a single cell (white).

In the practical implementation, condensed corner and curve cells have their geometric information stored in three arrays, for all possible cell pairs (ξ_i,ξ_j) in Ξ^{BCo} and Ξ^{BCu}, i.e., three arrays for corner pairs and another three for curve pairs, where *i* and *j* denote different coordinates (*x*, *y*):

1. $B^t_{i,j}$ for storing angle compatibility between pairs of corners or pairs of curves, i.e., if they have similar orientations within a specified margin as shown in Eq. 16 for corners and in Eq. 17 for curves;

2. $A^t_{i,j}$ for storing angles between cell pairs, see Eq. 18; and

3. $D^t_{i,j}$ for distances between cell pairs, see Eq. 19.

Mathematically, they are assigned as follows:

$$B^{Corner}_{i,j}=\begin{cases}\text{ON} & \text{if } \left[\hat{\phi}(\lambda_i)\pm\frac{1}{2}\breve{\phi}(\lambda_i)\right]\wedge\left[\hat{\phi}(\lambda_j)\pm\frac{1}{2}\breve{\phi}(\lambda_j)\pm\Delta_e\right]\neq\phi\\ \text{OFF} & \text{otherwise.}\end{cases} \tag{16}$$

$$B^{Curve}_{i,j}=\begin{cases}\text{ON} & \text{if connected by active }\Omega\text{ cells}\\ \text{OFF} & \text{otherwise.}\end{cases} \tag{17}$$

$$A^t_{i,j}=180^{\circ}/\pi\cdot\arctan\left((y_j-y_i)/(x_j-x_i)\right) \tag{18}$$

$$D^t_{i,j}=\max\left(\left|x_j-x_i\right|,\left|y_j-y_i\right|\right) \tag{19}$$

Finally, array $B^{Corner}_{i,j}$ is also checked for main-diagonal symmetry, such that only bidirectionally connected features will remain,

$$B^{Corner}_{i,j}=B^{Corner}_{i,j}\bigcap B^{Corner}_{j,i} \tag{20}$$

3 Final shape retrieval

Below, the shape repertoire is denoted by {*S*, *R*, *T*, *C*, *E*}: square, rectangle, triangle, circle and ellipse. Trapeziums and parallelograms are also detected by using the rules which apply to rectangles. All possible combinations coded in the Ξ layer are processed, and candidate shapes are validated using the following rules: the correct number of features, their relative distances, connectivity and internal angles, and the centre of the shape. Specifically:

Trapeziums and parallelograms are also detected by using the rules which apply to rectangles. All possible combinations coded in the Ξ layer are processed, and candidate shapes are validated using the following rules: the correct number of features, their relative distances, connectivity and internal angles, and the centre of the shape. Specifically:

1. A candidate shape must possess a **correct number of features** of the type corner or curve which match the shape model. A square or rectangle has to include four condensed corner cells *a* to *d*. In the case of a triangle there are three, *a*, *b* and *c*. Mathematically,

$$\exists \xi_i = \text{Corner}, \forall i \in \{a,b,c,d\} \tag{21}$$

$$\forall \{S,R\},\ B_{i,j} = \text{ON},\ \forall i,j \in \{a,b,c,d\}, i \neq j \tag{22}$$

$$\forall T, B_{i,j} = \text{ON},\ \forall i,j \in \{a,b,c\}, i \neq j. \tag{23}$$

In the case of a circle or ellipse, there must be three curve cells, *e*, *f* and *g*:

$$\exists \xi_i = \text{Curve}, \forall i \in \{e,f,g\} \tag{24}$$

$$\forall \{C,E\},\ B_{i,j} = \text{ON},\ \forall i,j \in \{e,f,g\}, i \neq j. \tag{25}$$

2. The **relative distances** between the shape's features must also match the shape model, i.e., a square must have four pairs of corners with about the same distances. The relative distances can be relaxed for detecting trapeziums and parallelograms. Similar but different processes are applied to the other shapes. It should be stressed that, in the particular cases of squares and rectangles, the distances are tested over adjacent corners, such that possible diagonals inside the shapes are inhibited:

$$\begin{aligned}\forall \{S,R\}, \min(D_{i,j}) > 0.5 \cdot \max(D_{i,j}), \forall i,j \in \{a,b,c,d\}, \\ i \neq j \wedge (i = a \wedge j \neq d) \wedge (i = b \wedge j \neq c)\end{aligned} \tag{26}$$

$$\forall T, \min(D_{i,j}) > 0.6 \cdot \max(D_{i,j}), \forall i,j \in \{a,b,c\}, i \neq j \tag{27}$$

$$\forall \{C,E\}, (D_{e,f} > 4) \wedge (D_{f,g} > 4) \wedge (D_{e,g} > 8), \forall i,j \in \{e,f,g\}, i \neq j. \tag{28}$$

3. The candidate shape must exhibit **connectivity** between shape features, especially between (condensed) corners, i.e., they must be linked by bar cells, with **confirmatory evidence** $CE(\xi_i, \xi_j)$ of connectivity, by analysing the number of bar cells with perpendicular orientations between corners or curves. Mathematically,

$$CE(\xi_i,\xi_j) = \begin{cases} \text{OFF} & \text{if } \#\xi_{(i \to j)} < 0.8 \cdot \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2} \\ & \quad \vee \#\xi_{(i \to j)} \left[= \text{Bar} \wedge \hat{\phi}(\xi_{(i \to j)}) \perp m \right] < 0.8 \\ \text{ON} & \text{otherwise,} \end{cases}$$

$$CE(\xi_i,\xi_j)=\begin{cases}\text{OFF} & \text{if}\#\ \xi_{(i\to j)} < 0.8\cdot\sqrt{(x_j-x_i)^2+(y_j-y_i)^2} \\ & \vee\#\ \xi_{(i\to j)}\left[=\text{Bar}\wedge\hat{\phi}(\xi_{[i\to j]})\perp m\right]<0.8 \\ \text{ON} & \text{otherwise,}\end{cases} \tag{29}$$

with **#** the number of cells and $m=(y_j-y_i)/(x_j-x_i)$.

4. For polygons, the sum of the **internal angles** of the candidate shape must match the model shape. Mathematically,

$$\forall\{S,R\},\sum A_{i,j}\approx 360^{\circ},\forall i,j\in\{a,b,c,d\},i\neq j \tag{30}$$

$$\forall T,\sum A_{i,j}\approx 180^{\circ},\forall i,j\in\{a,b,c\},i\neq j. \tag{31}$$

5. For circles and ellipses, a **centre of the shape** is estimated using the intersections of lines perpendicular to tangents of curve cells. The intersection point of two perpendicular lines yields an estimate of the shape's centre (x_c,y_c),

$$x_c=(y_c-y_{m2}+m_2\cdot x_{m2})/m_2;\ \mathrm{y}_c=m_2(x_c-x_{m1})+y_{m1}, \tag{32}$$

With

$m_1=(x_e-x_f)(y_f-y_e)$, $m_2=(x_f-x_g)(y_g-y_f)$, $x_{m_1}=(x_e+x_f)/2$, $x_{m_2}=(x_f+x_g)/2$, $y_{m_1}=(y_e+y_f)/2$, and $y_{m_2}=(y_f+y_g)/2$.

This process is applied to all triplets of curve cells in 9×9 neighbourhoods of the Ξ layer. Resulting intersection points are then averaged to obtain a single centre estimate. Of course, this solution is less accurate in the case of ellipses.

In summary, specific shapes are detected by activating detection cells which apply the rules explained above: a square and rectangle have to obey the activation rules of Eqns 21, 22, 26, 29 and 30, all at the same time. For a triangle these are Eqns 21, 23, 27, 29 and 31, and for a circle and ellipse Eqns 24, 25, 28, 29 and 32 apply.

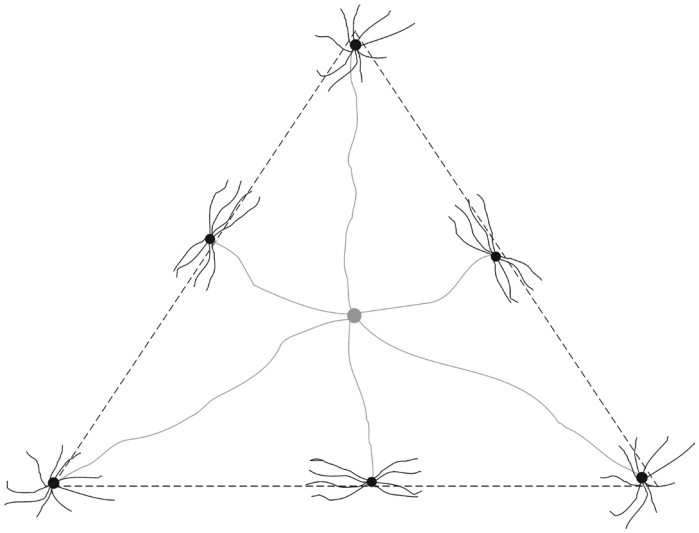

Fig. 7. A biologically plausible model for detecting a triangle may employ a grouping cell with long horizontal dendrites. These connect to other grouping cells with small dendritic fields for checking detected corners and straight bars. Clusters of cells are required to detect triangles with different rotations and sizes, and such clusters must exist at many positions.

The practical implementation of final shape retrieval is based on a full and serial search of the feature maps, applying the criteria as described above. A biologically plausible and fast implementation must involve a parallel search. For low- and mid-level geometry, small neighbourhoods are processed, from 3×3 up to 9×9 for connectivity and curvature (Fig. 4). All these local processes can be modelled by cells with small dendritic fields. Obviously, this is not possible for final shape retrieval, as the shapes can have different sizes and rotations. Hence, one must assume clusters of grouping cells with long horizontal dendrites.

For example, in the case of an equilateral triangle we can assume a cell with its body close to the centre of the triangle; see Fig. 7. It needs three horizontal dendrites, at 60º, of about the same length, which each connect at the end to another cell with a small dendritic field for checking whether a corner has been detected there. It also needs three horizontal dendrites, between the other three and ending between the corners, for checking whether straight bars have been detected there. This can be done by connecting them to cells with elongated dendritic fields, about perpendicular to the horizontal dendrites. We can assume that all grouping cells with the small dendritic fields exist at all positions and with many rotations. Therefore, they can be employed for other shapes. The only new requirement is the grouping cell with the right horizontal dendrites which connect to the right other cells. If the cell responds, its body or axon implicitly codes the triangle and its position.

The above example is a simplification, because the triangle can have a different position, rotation and size. One must therefore assume that clusters of triangle cells exist, at each position, for dealing with rotation and size. The same applies to clusters devoted to the other shapes, but also to e.g. triangles which are not equilateral. Partial occlusions are a further complication. Although the model can deal with partially occluded bars (see Discussion), it cannot yet deal with occluded corners. All this is subject to ongoing research.

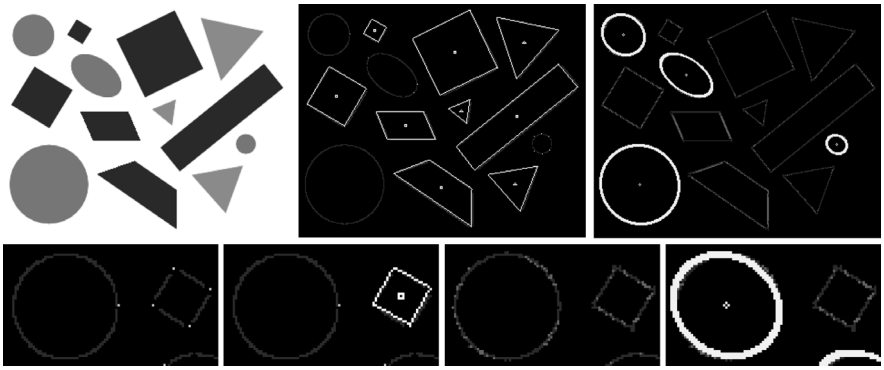

Fig. 8. Top row, from left: artificial test image with different shapes, rotations and sizes, and detected shapes. Bottom row: magnification of the top-left corner of the test image, showing from left to right corner/bar detection with condensed corners, square detected, curve/bar detection, and circle detected.

For a better comprehension of the results, Fig. 8 (top-left) shows an artificial test image with different squares, rectangles, trapeziums, triangles, circles and ellipses, with different rotations and sizes. Detected shapes and their centres are shown to the right. The bottom row shows a magnification of the top-left corner of the test image with, left to right: corner/bar detection with condensed corners, the square detected, curve/bar detection, and the circle detected.

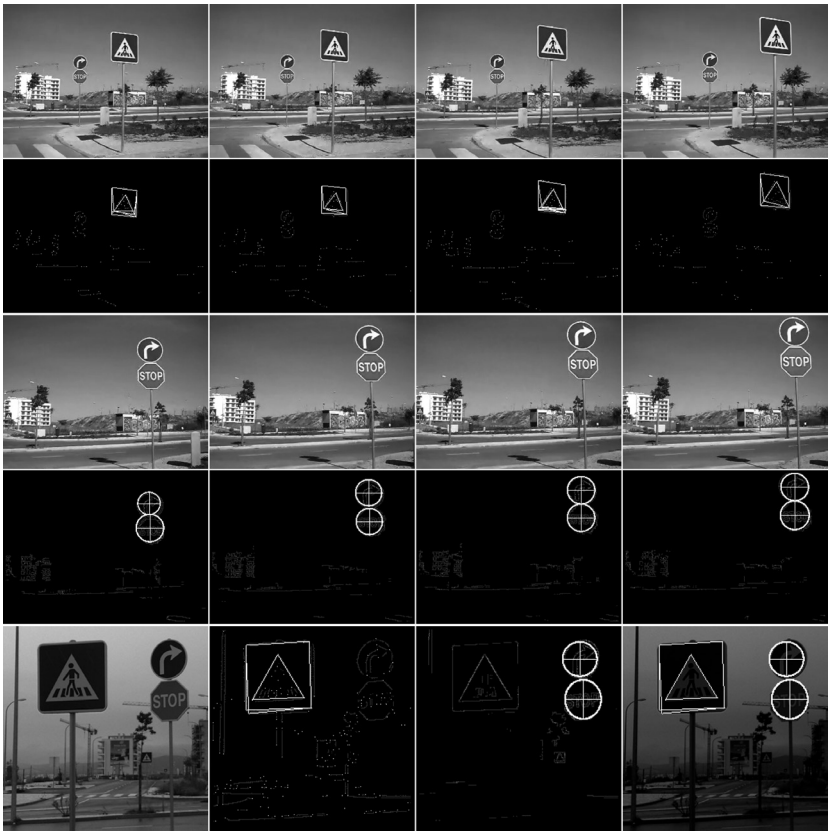

Fig. 9. Top four rows: two series of frames of a sequence acquired from a moving car with detected squares and triangles (2nd row) and circles and ellipses (4th row). Bottom row, left to right: another input image with traffic signs, detected triangles and squares, detected circles and ellipses, and all detected shapes superimposed on the input image.

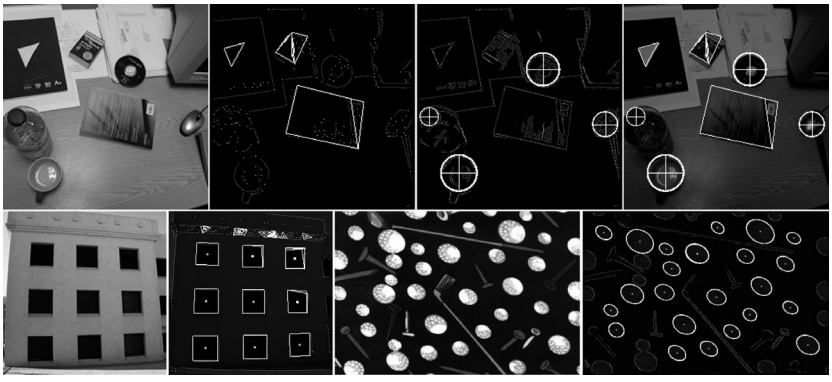

Fig. 10. Top, from left: image of an office desk, detected triangles and squares, detected circles and ellipses, and all detected shapes superimposed on the input image. Bottom: results in the case of a building and golf equipment.

The top and third rows in Fig. 9 show two parts, of four frames each, of a sequence acquired from a moving car. The second row shows squares and triangles detected in the frames on the top row, whereas the fourth row shows circles and ellipses detected in the frames on the third row. The bottom row shows (from left to right) a frame with more traffic signs from another sequence and all detected shapes superimposed on the input image. All frames were resized to about 256×256 pixels, and the processing time of each was about 0.2 seconds on a normal PC. Figure 10 shows more results in the case of an office desk (top row), a building, and golf equipment, with successfully detected squares, trapeziums, triangles, circles and ellipses. In the case of the golf equipment, we can see that some balls have not been detected; these were too close to other objects such that their edges were not well separated.

As can be seen in Figs 9 and 10, most important shapes as defined by colour contrast – colour conspicuity, the only information exploited here – have been detected, providing local gist of segregated objects in a spatial layout map, which can be used for subsequent object recognition by a sequential process steered by Focus-of-Attention.

4. Discussion

Multi-scale representations of lines, edges and keypoints, extracted on the basis of simple, complex and end-stopped cells in cortical areas V1 and V2, can be used for invariant object categorisation and recognition (Rodrigues & du Buf, 2009a,b). These representations are complemented by saliency maps of colour, texture, disparity and motion information, which are thought to play an important role in Focus-of-Attention or FoA (Elazary & Itti, 2008; Martins et al., 2009). This processing is done in the normal pathway with ventral and dorsal (what and where) data streams, which proceed from the LGN via V1 etc. to the prefrontal cortex. These data streams are bottom-up but with top-down attentional modulation from the prefrontal cortex down to the LGN (Saalmanna & Kastner, 2009).

As postulated by Rensink (2000), there may be two other subsystems for gist vision and spatial layout. These must be very fast, because they (1) serve to bias specific data paths related to specific objects in memory, i.e., the context serves to pre-select typical objects such that all objects held in memory which are out-of-context can be ignored, and (2) they prepare FoA for directing attention and our eyes to regions where important objects are expected. Global scene gist, for which computational models have already been developed (Bar, 2004; Siagian & Itti, 2007; Ross & Oliva, 2010; Rodrigues & du Buf, 2011), cannot be directly linked to spatial layout, because the latter implies, by definition, a localised analysis: which types of objects are about where in a scene. The missing link can

consist of local object gist, even with the possibility that this is extracted before global scene gist and, once local gist is available, global gist can be extracted completely at semantic level: detected objects determine the context and scene. In addition, object gist may contribute to solving the object segregation problem, i.e., if objects are complex in terms of coloured and textured regions. The class of general objects remains subject to further research, but here we have shown that at least *elementary* shapes indicating many man-made objects can be dealt with. The model explored in this paper only exploits local colour contrast or colour conspicuity. As explained in the Introduction, apart from the standard retinal ganglion cells there also are non-standard cells which code edges and their motion (Fig. 1). Hence, colour information could be combined with texture and motion information for developing a retinal model of local object gist, retinal meaning the use of retinal information at some higher level, with the possibility that also disparity information could be integrated. The lowest level where this could happen is in the LGN, after the optic chiasm where information from the left and right eyes can be combi

Fig. 11. Partial occlusions. Top: because of colour contrast, only the inner circle and triangle of the traffic signs have been detected. Bottom: two partially occluded monitors in a laboratory scene.

The model as developed employs the normal processing strategy in the brain, with massively parallel processing at a low level and increasing complexity at a higher level. Using few cell layers it is possible to extract strictly local syntactical

features and to combine them into "local-global" features like bars and corners, after which global semantics of elementary shapes can be extracted: man-made objects, often with square, rectangular, trapezoidal, triangular, circular and elliptical shapes. In addition, it is rather trivial to extract also the centres of these shapes. What is *not* trivial is to extract the shapes and their correct centres when they are partially occluded. In the current model, occlusions are already possible but only if the necessary corners of a shape are visible. Examples of detected shapes with such occlusions can be seen in Fig. 11.

The shapes' rules as specified and applied in this paper must be complemented for dealing with other partial occlusions. This can be done with relaxation rules, such that a rectangle with one occluded corner can nevertheless be detected as a rectangle and not as a triangle. In such a case, more emphasis should be on the rectangle's edges and their parts, if some of the edges are also partially occluded. The latter is a direct application of Gestalt Theory's rules of proximity and good continuation. The rule concerning convexity has already been implemented, though implicitly, because all shapes in the repertoire are convex, whereas the symmetry rule can be applied to parallel edges in case of squares and rectangles. The most interesting question concerns the way in which local object gist and the non-standard retinal cells can be integrated in the normal pathway for invariant object recognition. The trivial part of the answer is that the spatial layout map – the centres of shapes and their type – can be exploited in the prefrontal cortex for (a) biasing all objects in memory with the same shapes, and (b) updating the FoA map in order to prepare saccadic eye movements. Much less trivial are the non-standard cells and the new pathways as discussed by Masland & Martin (2007). It is possible that not only top-down attentional modulation from the prefrontal cortex influences processing down to the lower levels V4, V2, V1 and even the LGN, but that the same occurs bottom-up and at the same time. The difference may be that top-down modulation can be a serial process whereas bottom-up modulation can be a parallel one. Such questions are very speculative, and it may take some years before we know more about these issues, both the non-standard cells and good computational models of them, and their pathways to and roles in other visual areas.

Acknowledgements

This research was supported by the Portuguese Foundation for Science and Technology (FCT) through the PIDDAC Program funds (ISR/IST plurianual funding), EC project NeuroDynamics (FP7-ICT-2009-6 PN: 270247), FCT project Blavigator (RIPD/ADA/109690/2009) and FCT PhD grant to Jaime A. Martins (SFRH-BD-44941-2008).

Summary

High-level vision is based on semantic representations of scenes or context and important objects therein. The bottom-up data streams, from retina via LGN to V1 and further, are devoted to moving objects and motor control in the dorsal where stream, and to invariant object recognition in the ventral what stream. They are steered, top-down, by attention and short-time memory in the ventral and dorsal regions of the prefrontal cortex. However, these processes are bootstrapped, and probably continuously guided, by an extremely fast analysis devoted to scene gist and spatial layout, i.e., which types of objects are about where in the scene. Most research has been devoted to global scene gist, but in this paper we present an alternative approach which addresses local object gist for simultaneous object segregation, attention, and spatial layout. The proposed model only exploits colour information, although texture, motion and disparity information can also be integrated. Specifically, we focus on man-made objects which are dominated by a simple shape repertoire: squares, rectangles, trapeziums, triangles, circles and ellipses. It is shown that such shapes can be detected by a hierarchy of a few cell layers, with strictly bottom-up or data-driven processing. We argue that this processing may occur in very early vision, possibly only employing signals from non-standard retinal ganglion cells. Although proposed to play a role in the fast dorsal stream, similar processing may occur in the slower ventral stream.

Keywords: Low-level vision, local gist, object segregation, shape extraction, spatial layout.

Zusammenfassung

Die visuelle Wahrnehmung unseres Umfelds (high-level vision) basiert auf der semantischen Repräsentation entweder ganzer Szenen oder spezieller Kontexte mit darin eingebetteten wichtigen Objekten. Der Datenfluß von der Retina über das LGN zu V1 und darüber hinaus fokussieren auf sich bewegende Objekte und motorische Kontrolle im dorsalen 'Wo'-Fluß, und auf invariante Objekterkennung im ventralen 'Was'-Fluß. Diese Prozesse werden durch Aufmerksamkeitsprozesse und Aktivitäten im Kurzzeitgedächtnis in ventralen und dorsalen Bereichen des präfrontalen Kortex gesteuert und gleichzeitig kontinuierlich durch extrem schnelle Analysen der 'essentiellen Struktur' (scene gist) und der räumlichen Anordnung kontrolliert, die sich auf den Ort verschiedener Typen von Objekten in der Szene richten. Die Forschung hat sich bisher meistens auf die 'globale essentielle Struktur' (global scene gist) konzentriert. In dieser Arbeit präsentieren wir einen alternativen Ansatz, der sich auf die lokale 'essentielle Objektstruktur' (object gist) für simultane Objekttrennung, Aufmerksamkeit und räumliche Anordnung richtet. Das vorgeschlagene Modell nimmt in erster Linie die Verarbeitung von Farbinformationen an, wobei Textur-, Bewegungs- und Disparitätsinformation ebenfalls integriert werden können. Wir konzentrieren uns speziell auf geometrische Objekte aus einem einfachen Repertoire von Formen: Quadrate, Rechtecke, Trapeze, Dreiecke, Kreise und Ellipsen. Es wird gezeigt, dass solche Gestalten durch eine Hierarchie von wenigen Zellschichten mit datengetriebener Verarbeitung (bottom-up) entdeckt werden kann. Es wird argumentiert, dass diese Prozesse in einem der ersten Stadien der Verarbeitung visueller Information ablaufen, wobei möglicherweise nur Informationen aus der Aktivität bestimmter retinaler Ganglionzellen verwendet werden. Obwohl vorgeschlagen wird, dass diese Prozesse in der schnellen, dorsalen Informationsverarbeitung ablaufen, kann angenommen werden, dass ähnliche Prozesse in der langsameren ventralen Informationsverarbeitung ablaufen.

Schlüsselwörter: Niedrige Vision, lokale Struktur, Objekttrennung, Formextraktion, räumliche Anordnung.

References

Bar, M. (2004): Visual objects in context, *Nature Reviews: Neuroscience 5*, 619–629.

Bar, M., Kassam, K., Ghuman, A., Boshyan, J., Schmid, A., Dale, A., Hämäläinen, M., Marinkovic, K., Schacter, D., Rosen, B. & Halgren, E. (2006): Top-down facilitation of visual recognition, *PNAS 103*, 449–454.

Biederman, I. & Kalocsai, P. (1997): Neurocomputational bases of object and face recognition, *Philosoph. Trans. Royal Society: Biol. Sciences 352*, 1203–1219.

Bomberger, N. & Schwartz, E. (2005): The structure of cortical hypercolumns: Receptive field scatter may enhance rather than degrade boundary contour representation in V1, *J. of Vision 5*, 891.

Craft, E., Schütze, H., Niebur, E. & von der Heyd, R. (2007): A neural model of figure-ground organization, *J. Neurophysiol 97*, 4310–4326.

Creem-Regehr, S. (2009): Sensory-motor and cognitive functions of the human posterior parietal cortex involved in manual actions, *Neurobiology of Learning and Memory 91*, 166–171.

Crick, F. & Koch, C. (2003): A framework for consciousness, *Nature Neuroscience 6*, 119–126.

Deco, G. & Rolls, E. (2005): Attention, short term memory, and action selection: a unifying theory, *Progress in Neurobiology 76,* 236–256.

Elazary, L. & Itti, L. (2008): Interesting objects are visually salient, *J. of Vision 8*, 1–15.

Farivar, R. (2009): Dorsal-ventral integration in object recognition, *Brain Research Reviews 61*, 144–153.

Faubel, C. & Schöner, G. (2009): A neuro-dynamic architecture for one shot learning of objects that uses both bottom-up recognition and top-down prediction, *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 3162–3169.

Finlayson, G., Schiele, B. & Crowley, J. (1998): Comprehensive colour image normalization, *Proc. 5th Europ. Conf. Computer Vision*, Berlin (Germany), Springer LNCS 1406, 475–490.

Greene, M. & Oliva, A. (2009): Recognition of natural scenes from global properties: Seeing the forest without representing the trees, *Cognitive Psychology 58*, 137–176.

Houtkamp, R. & Roelfsema, P. (2010): Parallel and serial grouping of image elements in visual perception, *J. of Experimental Psychology: Human Perception and Performance 36*, 1443–1459.

Hubel, D. (1995): *Eye, Brain and Vision*, Scientific American Library, Vol. 22, New York.

Karttunen, R. & Häkkinen, S. (1982): Discrimination of traffic signs in the peripheral areas of the field of vision, *Reports from Liikenneturva 25/1982*, Central Organization for Traffic Safety, Helsinki (Finland).

Kastner, S., Schneider, K. & Wunderlich, K. (2006): Comparison of shape encoding in primate dorsal and ventral visual pathways, Chapter 8: *Beyond a relay nucleus: neuroimaging views on the human LGN*, Martinez-Conde, S., Macknik, S., Martinez, L., Alonso, J.-M. & Tse, P. (Eds), Progress in Brain Research, Elsevier.

Kiani, R., Esteki, H., Mirpour, K. & Tanaka, K. (2007): Object category structure in response patterns of neuronal population in monkey inferior temporal cortex, *J. Neurophysiology 97*, 4296–4309.

Kohn, A. & Movshon, J. (2003): Neuronal adaptation to visual motion in area MT of the macaque, *Neuron 39*, 681–691.

Konen, C. & Kastner, S. (2008): Two hierarchically organized neural systems for object information in human visual cortex, *Nature Neuroscience 11*, 224–231.

Krüger, N., Lappe, M. & Wörgötter F. (2003): Biologically motivated multi-modal processing of visual primitives, *Proc. Symp. Biologically-Inspired Machine Vision, Theory and Application*, 53–59.

Lehky, R. & Sereno, A. (2007): Comparison of shape encoding in primate dorsal and ventral visual pathways, *J. of Neurophysiology 97*, 307–319.

Luoma, J. (1992): Immediate responses to road signs of alerted and unaltered drivers: An evaluation of the validity of eye movement method, *Transportation Research Board annual meeting*, Jan. 1992, Washington DC.

Martens, M. & Fox, M. (2007): Does road familiarity change eye fixations? A comparison between watching a video and real driving, *Transportation Research Part F 10*, 33–47.

Martins, J.A., Rodrigues, J. & du Buf, J.M.H. (2009): Focus of attention and region segregation by low-level geometry, Proc. Int. Conf. on Computer Vision - Theory and Applications, Lisbon, Portugal 2 (2009) 267–272.

Masland, R. & Martin, P. (2007): The unsolved mystery of vision, *Current Biology 17*, R577–82.
Oliva, A. & Torralba, A. (2006): Building the gist of a scene: the role of global image features in recognition, *Progress in Brain Res.: Visual Perception 155*, 23–26.
Peng, Q. & Shi, B. (2010): The changing disparity energy model, *Vision Research 50*, 181–192.
Pinna, B. & Reeves, A. (2006): Lighting, backlighting and watercolor illusions and the laws of figurality, *Spatial Vision 19*, 341–373.
Pinna, B. (2010): New gestalt principles of perceptual organization: an extension from grouping to shape and meaning. *Gestalt Theory 32*, 11–78.
Pugeault, N., Wörgötter, F. & Krüger, N. (2010): Disambiguating multi-modal scene representations using perceptual grouping constraints, *PLoS ONE 5*, 16.
Rensink, R. (2000): The dynamic representation of scenes, *Visual Cognition 7*, 17–42.
Rodrigues, J. & du Buf, J. (2011): A cortical framework for scene categorization, *Proc. Int. Conf. on Computer Vision – Theory and Applications*, Vilamoura, Portugal, 5-7 March (2011), 364–371.
Rodrigues, J. & du Buf, J.M.H. (2009a): Multi-scale lines and edges in V1 and beyond: Brightness, object categorization and recognition, and consciousness, *BioSystems 95*, 2206–2260.
Rodrigues, J. & du Buf, J.M.H. (2009b): A cortical framework for invariant object categorization and recognition, *Cognitive Processing 10*, 243–261.
Ross, M. & Oliva, A. (2010): Estimating perception of scene layout properties from global image features, *J. of Vision 10*, 1–25.
Rubin, E. (1921): *Visuell wahrgenommene Figuren*, Kobenhavn: Gyldendalske.
Saalmanna, Y. & Kastner, S. (2009): Gain control in the visual thalamus during perception and cognition, *Current Opinion in Neurobiology 19*, 408–414.
Siagian, C. & Itti, L. (2007): Rapid biologically-inspired scene classification using features shared with visual attention, *IEEE Tr. on Robotics 29*, 300–312.
Smith, A., Wall, M., Williams, A. & Singh, K. (2006): Sensitivity to optic flow in human cortical areas MT and MST, *European J. of Neuroscience 23*, 561–569.
Stoerig, P. & Cowey, A. (1997): Blind sight in man and monkey, *Brain 120*, 535–559.
Troncoso, X., Macknik, S. & Martinez-Conde, S. (2011): Vision's first steps: Anatomy, physiology, and perception in the retina, lateral geniculate nucleus, and early visual cortical areas, Chapter 2, *Visual Prosthetics: Physiology, Bioengineering, Rehabilitation*, Dagnelie, G. (Ed.), Springer Science+Business Media, 23–57.
Wall, M. & Smith, A. (2008): The representation of egomotion in the human brain, *Current Biology 18*, 191–194.
Wertheimer, M. (1923): Untersuchungen zur Lehre von der Gestalt. *Psychologische Forschung 4(II)*, 301–350.

Postal address of all authors
Vision Laboratory, Institute for Systems and Robotics (ISR), University of the Algarve (FCT and ISE), 8005-139 Faro, Portugal

Jaime Afonso Martins, b. 1980, graduated 2001 in Electronics and Telecommunications Engineering and 2007 in Clinical Psychology at the University of the Algarve. Supported by a research grant from the Portuguese Foundation for Science and Technology, he is since 2008 pursuing a PhD degree in the Vision Laboratory at UAlg. His major interests concern areas which link cognitive neuroscience to computer and human vision.
E-mail: jamartins@ualg.pt

João Rodrigues, b. 1971, graduated in Electrical Engineering in 1993 at the University of Trás-os- Montes and Alto Douro (Portugal). He obtained his PhD degree in 2008 at the University of the Algarve, where he lectures computer science courses. Since 1996 senior researcher of the Vision Laboratory (UAlg), member of the Institute for Systems and Robotics (Lisbon). His major research interests concern human vision: gist, attention and object categorisation and recognition.
E-mail: jrodrig@ualg.pt

Hans du Buf, b. 1951, received a PhD degree at the Technical University of Eindhoven (The Netherlands) in 1987. He then worked at the Swiss Federal Institute of Technology in Lausanne (Switzerland). 1994 associate professor at the Dept. of Electronics and Computer Science, University of the Algarve. He conducts research

in image processing, pattern recognition, computer graphics and visual perception. He is a member of the Portuguese chapter of the IAPR, and is associate editor of the International Journal of Pattern Recognition and Artificial Intelligence.
E-mail: dubuf@ualg.pt