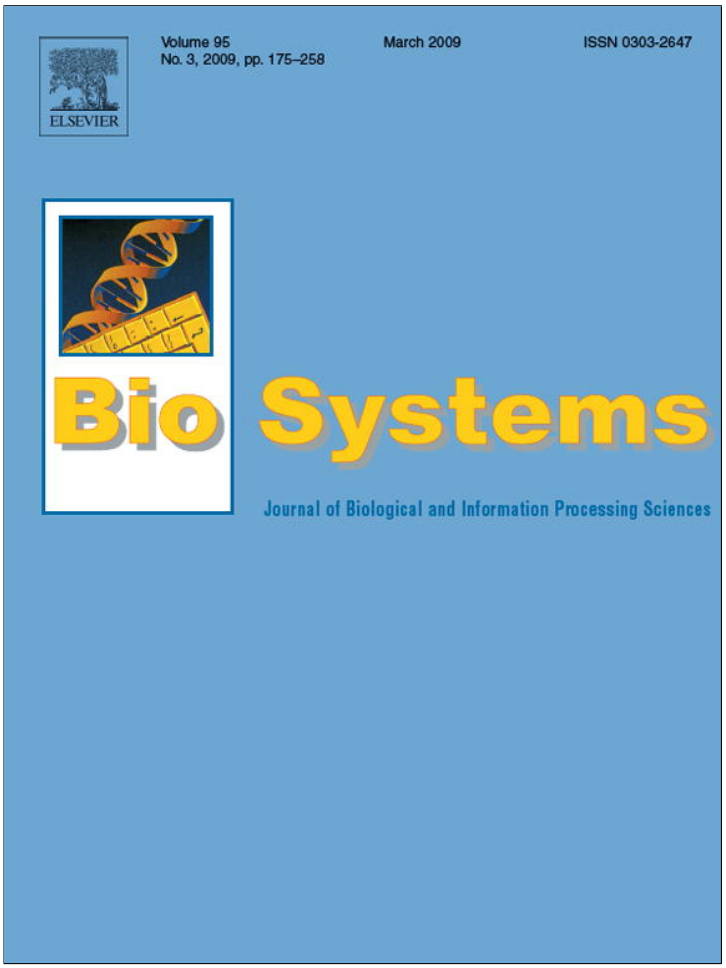

BioSystems 95 (2009) 206–226

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

Multi-scale lines and edges in V1 and beyond: Brightness, object categorization and recognition, and consciousness

João Rodrigues*, J.M. Hans du Buf

Vision Laboratory, Institute for Systems and Robotics, University of the Algarve, Campus de Gambelas – FCT, 8000-810 Faro, Portugal

ARTICLE INFO

Article history:
Received 4 October 2007
Received in revised form
19 September 2008
Accepted 22 October 2008

Keywords:
Visual cortex
Line/edge
Multi-scale
Reconstruction
Brightness
Segregation
Categorization
Recognition
Consciousness

ABSTRACT

In this paper we present an improved model for line and edge detection in cortical area V1. This model is based on responses of simple and complex cells, and it is multi-scale with no free parameters. We illustrate the use of the multi-scale line/edge representation in different processes: visual reconstruction or brightness perception, automatic scale selection and object segregation. A two-level object categorization scenario is tested in which pre-categorization is based on coarse scales only and final categorization on coarse plus fine scales. We also present a multi-scale object and face recognition model. Processing schemes are discussed in the framework of a complete cortical architecture. The fact that brightness perception and object recognition may be based on the same symbolic image representation is an indication that the entire (visual) cortex is involved in consciousness.

1. Introduction

The visual cortex detects and recognizes objects by means of the ventral "what" and dorsal "where" subsystems. The "bandwidth" of these systems is limited: only one object can be attended at any time (Rensink, 2000). In a current model by Deco and Rolls (2004) the ventral what system receives input from area V1 which proceeds through V2 and V4 to IT (inferior temporal cortex). The dorsal where system connects V1 and V2 through MT (medial temporal) to area PP (posterior parietal). Both systems are controlled, top-down, by attention and short-term memory with object representations in PF (prefrontal) cortex, i.e., a what component from PF46v to IT and a where component from PF46d to PP. The bottom-up (visual input code) and top-down (expected object and position) data streams are necessary for obtaining size, rotation and translation invariance, assuming that object views are normalized in visual memory.

Signal propagation from the retinas through the LGN (lateral geniculate nucleus) and areas V1, V2 etc., including feature extractions in V1 and groupings in higher areas, takes time. Object recognition is achieved in 150–200 ms and category-specific activation of PF cortex starts after about 100 ms (Bar, 2004). In addition, IT cortex first receives coarse-scale information and later fine-scale information. Apparently, one very brief glance is sufficient for the system to develop a gist of the contents from an image (Oliva and Torralba, 2006). This implies that some information propagates very rapidly and directly to "attention" in PF cortex in order to pre-select possible object-group templates and positions that then propagate down the what and where systems. This process we call object categorization, which cannot be obtained by the CBF (Categorical Basis Functions) model by Riesenhuber and Poggio (2000) because categorization (e.g. a cat) is obtained by grouping outputs of identification cells (cat-1, cat-2, cat-3). In other words, categorization would be obtained *after* recognition. In contrast, the LF (Low Frequency) model (Oliva et al., 2003; Bar, 2004) assumes that categorization is obtained *before* recognition: low-frequency information that passes directly from V1/V2 to PF cortex, although the LF information actually proposed consists of lowpass-filtered images, but not of e.g. outputs of simple and complex cells in V1 which are tuned to low spatial frequencies. The latter option will be explored in this paper.

After object categorization on the basis of coarse-scale information has narrowed the set of objects to be tested, the recognition process can start by also applying fine-scale information. We will focus on how such processes can be embedded in the architecture referred to above, with special focus on face recognition. Despite the impressive number and variety of computer-vision

* Corresponding author at: University of the Algarve, Escola Superior de Tecnologia, Campus da Penha, 8005-132 Faro, Portugal. Tel.: +351 289800100; fax: +351 289888405.
E-mail addresses: jrodrig@ualg.pt (J. Rodrigues), dubuf@ualg.pt (J.M.H. du Buf).

0303-2647/$ – see front matter © 2008 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2008.10.006

methods devised for faces and facial landmarks, see e.g. Yang et al. (2002), we show that very promising results with a cortical model can be obtained, even in the case of some classical complications involving changes of pose (frontal vs. 3/4), facial expression, some lighting and noise conditions, and the wearing of spectacles.

In computer vision there exists a vast literature, from basic feature extraction to object segregation, categorization and recognition, and from image reconstruction (coding and decoding) to scale stabilization to disparity, but much less in biological vision. We therefore continue with a very brief summary of approaches related to this paper, with special focus on biological methods.

In addition to a few general overviews, see e.g. Hubel (1995), Bruce et al. (2000), Rasche (2005) and Miikkulainen et al. (2005), there also are detailed and quantitative models of simple and complex cells (Heitger et al., 1992; Petkov and Kruizinga, 1997), plus various models for inhibitions (Heitger et al., 1992; Petkov et al., 1993b; Barth et al., 1998; Rodrigues and du Buf, 2006a), edge detection (Smith and Brady, 1997; Elder and Zucker, 1998; Kovesi., 1999; Grigorescu et al., 2003) and combined line and edge detection (Verbeek and van Vliet, 1992; van Deemter and du Buf, 2000; Rodrigues and du Buf, 2004, 2006a). Other models address saliency maps and Focus-of-Attention (Itti and Koch, 2001; Parkhurst et al., 2002; Deco and Rolls, 2004; Rodrigues and du Buf, 2006d), figure-ground segregation (Heitger and von der Heydt, 1993; Hupe et al., 2001; Zhaoping, 2003; Rodrigues and du Buf, 2006a) and object categorization (Riesenhuber and Poggio, 2000; Leibe and Schiele, 2003; Csurka et al., 2004; Rodrigues and du Buf, 2006a). Concerning faces, various approaches have been proposed, from detecting faces and facial landmarks to the influence of different factors such as race, gender and age (Delorme and Thorpe, 2001; Yang et al., 2002; Ban et al., 2003; Rodrigues and du Buf, 2005b), including final face recognition (Kruizinga and Petkov, 1995; Zhao et al., 2003; Rodrigues and du Buf, 2006c, d). Yet other models have been devised for disparity (Fleet et al., 1991; Ohzawa et al., 1997; Qian, 1997; Rodrigues and du Buf, 2004), automatic scale selection (Lindeberg, 1994), visual reconstruction (Rodrigues and du Buf, 2006b) and brightness perception (du Buf, 2001). In this paper we show that one basic process, namely line and edge detection in V1 (and possibly V2), can be linked to most if not all the topics mentioned above, even to consciousness. We present an improved scheme for multi-scale line/edge extraction in V1, which is truly multi-scale with no free parameters. We illustrate the line/edge interpretation (coding and representation) for automatic scale selection and explore the importance of this interpretation in object reconstruction, segregation, categorization and recognition. Since experiments with possible Low-Frequency models based on lowpass-filtered images, following Bar (2004), gave rather disappointing results, which is due to smeared blobs of objects that lack any structure, we propose that categorization is based on coarse-scale line/edge coding, and that recognition involves all scales. Processing schemes are discussed in the framework of a complete cortical architecture. We emphasize that the multi-scale keypoint information also extracted in V1, which was shown to be very important for detection of facial landmarks and entire faces (Rodrigues and du Buf, 2006d), and other important features such as texture information that can be retrieved from bar and grating cells (du Buf, 2007), will not be employed here, because we want to focus completely on the multi-scale line/edge information in V1 and beyond. Therefore, this paper complements the previous one dedicated to keypoints (Rodrigues and du Buf, 2006d).

In Section 2 we present line/edge detection and classification in single- and multi-scale contexts, plus the application of non-classical receptive field (NCRF) inhibition. Section 3 illustrates the visual reconstruction model in relation to brightness perception. Section 4 deals with object segregation, Section 5 with automatic scale selection, and Section 6 with object categorization. This is followed by face recognition in Section 7 and consciousness in Section 8. We conclude with a final discussion in Section 9.

2. Line and Edge Detection and Classification

In many models it is assumed that Gabor quadrature filters provide a good model of receptive fields (RFs) of cortical simple cells. In the spatial domain (x, y) they consist of a real cosine and an imaginary sine, both with a Gaussian envelope (Lee, 1996; Grigorescu et al., 2003; Rodrigues and du Buf, 2006d). As in Rodrigues and du Buf (2006d), an RF is given by

$$G_{\lambda,\sigma,\theta,\varphi}(x,y) = \exp\left(-\frac{\tilde{x}^2+\gamma\tilde{y}^2}{2\sigma^2}\right)\cdot\cos(2\pi\frac{\tilde{x}}{\lambda}+\varphi), \quad (1)$$

with $\tilde{x} = x\cos\theta + y\sin\theta$ and $\tilde{y} = y\cos\theta - x\sin\theta$, where $1/\lambda$ is the spatial frequency, λ being the wavelength. Here we apply exactly the same parameter values. For the bandwidth σ/λ we use 0.56, which yields a half-response width of one octave (σ determines the size of the RF). The angle θ determines the orientation (we use 8 orientations), and φ the phase symmetry (0 or $-\pi/2$). We apply filters with an aspect ratio of $\gamma = 0.5$. Below, the scale s of analysis will be given in terms of λ expressed in pixels, where $\lambda = 1$ corresponds to 1 pixel. Most images shown in this paper have a size of 256×256 pixels. We can apply a linear scaling between $f_{\min}$ and $f_{\max}$ with either a few discrete scales or hundreds of almost contiguous scales. Responses of even and odd simple cells, which correspond to the real and imaginary parts of a Gabor filter, are denoted by $R^E_{s,i}(x,y)$ and $R^O_{s,i}(x,y)$, s being the scale, i the orientation ($\theta_i = i\pi/N_\theta$) and N_θ the number of orientations (we use $N_\theta = 8$). Responses of complex cells are modeled by the modulus following $C_{s,i}(x,y) = [\{R^E_{s,i}(x,y)\}^2 + \{R^O_{s,i}(x,y)\}^2]^{1/2}$. A basic scheme for single-scale line and edge detection based on responses of simple cells works as follows (van Deemter and du Buf, 2000): a positive (negative) line is detected where R^E shows a local maximum (minimum) and R^O shows a zero crossing. In the case of edges the even and odd responses are swapped. This gives 4 possibilities for positive and negative events. An improved scheme (Rodrigues and du Buf, 2004) consists of combining responses of simple and complex cells: simple cells serve to detect positions and event types, whereas complex cells are used to increase the confidence. Since the use of Gabor modulus (complex cells) implies a loss of precision at vertices (du Buf, 1993), increased precision was obtained by considering multiple scales (i.e. a few neighboring micro-scales).

The algorithms described above work reasonably well but there remain a few problems: (a) either one scale is used or only a very few scales for increasing confidence, (b) some parameters must be optimized for specific input images or even as a function of scale, (c) detection precision can still be improved, and (d) detection continuity at curved lines/edges must be guaranteed. Therefore we present an improved algorithm with no free parameters, truly multi-scale and with new solutions for problems (c) and (d).

With respect to precision, simple and complex cells respond beyond line and edge terminations, for example beyond the corners of a rectangle. In addition, at line or edge crossings and junctions, detection leads to continuity of the dominant events with biggest amplitudes but to gaps in the sub-dominant events. These gaps must be reduced in order to reconstruct continuity. Both problems can be solved by introducing new inhibition schemes, like the radial and tangential ones used in the case of keypoint operators (Rodrigues and du Buf, 2006d). Here we use lateral (L) and

cross-orientation (C) inhibition, defined as

$$I^L_{s,i}(x,y) = [C_{s,i}(x + dC_i, y + dS_i) - C_{s,i}(x - dC_i, y - dS_i)]^+ + [C_{s,i}(x - dC_i, y - dS_i) - C_{s,i}(x + dC_i, y + dS_i)]^+; \quad (2)$$

$$I^C_{s,i}(x,y) = [C_{s,(i+N_\theta/2)}(x + 2dC_i, y + 2dS_i) - 2.C_{s,i}(x,y) + C_{s,(i+N_\theta/2)}(x - 2dC_i, y - 2dS_i)]^+, \quad (3)$$

where $(i + N_\theta/2) \perp i$, with $C_i = \cos\theta_i$ and $S_i = \sin\theta_i$, $d = 0.6$ s, and $[\cdot]^+$ denotes halfwave rectification to suppress negative responses. Inhibition is applied to the responses of complex cells, where β controls the strength of inhibition (we always use $\beta = 1.0$):

$$\hat{C}_{s,i} = [C_{s,i}(x,y) - \beta(I^L_{s,i}(x,y) + I^C_{s,i}(x,y))]^+. \quad (4)$$

Fig. 1 shows a cross formed by two bars (top-left), the summation of L and C inhibition at $\theta = \{0, \pi/2\}$ (bottom) and the detection result (top-right) with no spurious events beyond the corners and with no gaps at the junctions. It should be emphasized that the same parameters ($d = 0.6$ s and $\beta = 1.0$) will be applied at all scales, with no need to optimize other parameters. Algorithms from computer vision (see below) are normally applied at one, fine scale and always need careful optimization for any input image.

Line and edge detection is achieved by assuming a few cell layers on top of simple and complex cells; see Fig. 2 for a wiring diagram. The first layer serves to select active regions and dominant orientations. At each position, responses of complex cells are summed, $\hat{C}_s = \sum_{i=0}^{N_\theta - 1} \hat{C}_{s,i}$, and if $\hat{C}_s > 0$ an output cell $A(x,y)$ is activated. At active output cells, the dominant orientation Id is selected by non-maximum suppression of $\hat{C}_{s,i}$. Since the processing is the same at all scales, we will drop scale-subscript s for clarity. Hence, $Id = \max_i \hat{C}_i$. In order to solve small inconsistencies at the pixel level, Id may be corrected by assigning the most frequent dominant orientation in the local neighborhood to the center position. This is done by counting the occurrences of all dominant orientations, and selecting the one with maximum count, in a window of size 3×3 pixels. The same window size is used at all scales.

In the second layer, event type and position are determined on the basis of active output cells (1st layer) and gated simple and complex cells. A first cell complex checks responses of simple cells R^E_{Id} and R^O_{Id} for a local maximum and a minimum, using a dendritic field size of $\pm\lambda/4$, λ being the wavelength of the simple cells (Gabor filter). Mathematically, using $\delta, \delta' \in [-\Delta, \Delta]$ with $\Delta = \lambda/4$ and $\delta \neq \delta'$, and assuming that $Id = 0$ such that the analysis is in x and the necessary rotation can be omitted, four "Boolean cells" MAX and MIN are defined:

$$\mathrm{MAX}^T_R(x) = \exists!\delta : \forall\delta' \quad R^T(x + \delta) > R^T(x + \delta') \quad (5)$$

and

$$\mathrm{MIN}^T_R(x) = \exists!\delta : \forall\delta' \quad R^T(x + \delta) < R^T(x + \delta'), \quad (6)$$

where type T is E (even) or O (odd). The outputs of the four cells are OR-ed and the active output cell $A(x,y)$ is inhibited if

$$\mathrm{MAX}(x,y) = \mathrm{MAX}^E_R(x,y) \vee \mathrm{MIN}^E_R(x,y) \vee \mathrm{MAX}^O_R(x,y) \vee \mathrm{MIN}^O_R(x,y) \quad (7)$$

is not true. A second cell complex does exactly the same on the basis of responses of complex cells, i.e.,

$$\mathrm{MAX}_{\hat{C}}(x) = \exists!\delta : \forall\delta' \quad \hat{C}(x + \delta) > \hat{C}(x + \delta'). \quad (8)$$

A third cell complex serves to detect zero-crossings in the responses of simple cells, again on $\pm\lambda/4$. Using $\varepsilon \ll \Delta$,

$$\mathrm{ZC}^T_R = \exists!\delta : R^T(x + \delta + \varepsilon) \cdot R^T(x + \delta - \varepsilon) < 0, \quad (9)$$

type T being E or O. If there is no zero-crossing, i.e., $\mathrm{ZC}_R(x) = \mathrm{ZC}^E_R(x) \vee \mathrm{ZC}^O_R(x)$ is false, the output cell A is inhibited. If there is a zero-crossing, only one of four event cells can be activated:

$$L^+(x) = \mathrm{MAX}^E_R(x) \wedge \mathrm{ZC}^O_R(x), \quad (10)$$

$$L^-(x) = \mathrm{MIN}^E_R(x) \wedge \mathrm{ZC}^O_R(x), \quad (11)$$

$$E^+(x) = \mathrm{MAX}^O_R(x) \wedge \mathrm{ZC}^E_R(x) \quad (12)$$

and

$$E^-(x) = \mathrm{MIN}^O_R(x) \wedge \mathrm{ZC}^E_R(x), \quad (13)$$

where $L^\pm$ and $E^\pm$ stand for positive and negative line and edge. Although explained in x assuming $Id = 0$, the same, but rotated processing is applied in case of any $Id \neq 0$, using (bi-linear) interpolation cells between positions of simple and complex cells.

In the third layer, the small loss of accuracy due to the use of responses of complex cells in the second layer is compensated. This is done by correcting local event continuity, considering the information available in the second layer, and allowing for curvature by taking into account responses of cells in the dominant orientation but also responses of cells tuned to two neighboring orientations in the neighborhood. The latter process is an extension of linear grouping (van Deemter and du Buf, 2000), a simplification of using banana wavelets (Krüger and Peters, 2000), and a mechanism for implementing Gestalt's rule of good continuity by means of local "association fields" (Field et al., 1993). This process is explained below. In the same layer, also event type and polarity are corrected by considering small neighborhoods, restoring continuity since cell responses may be distorted by interference effects when two events are very close (du Buf, 1993). As for correction of local dominant orientation at the pixel level in the first layer, this is achieved by considering neighborhoods of 3×3 pixels at all scales, but by counting the occurrences of the four event types $L^\pm$ and $E^\pm$ and assigning the most frequent event type to the center position.

The second row in Fig. 3 shows detection results at the finest scale $\lambda = 4$, with positive and negative lines and edges coded by different levels of gray (white, light gray, dark gray and black,

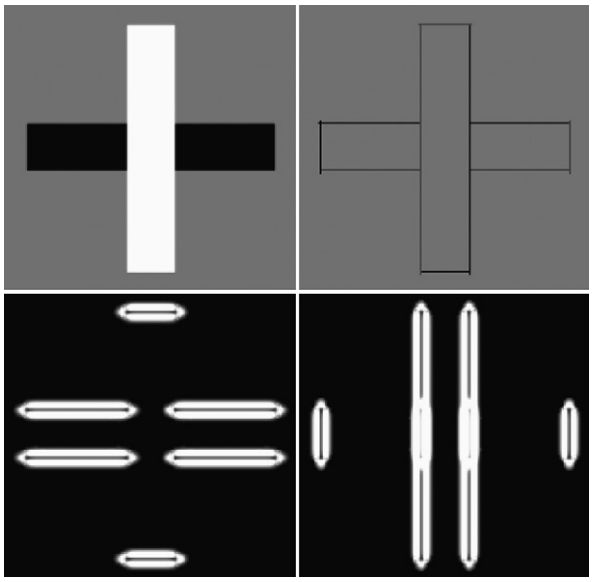

Fig. 1. Input pattern (top-left), the summation of lateral and cross-orientation inhibition for $\theta = \{0, \pi/2\}$ (bottom) and the detection result (top-right) with no spurious events beyond the corners and no gaps at the junctions.

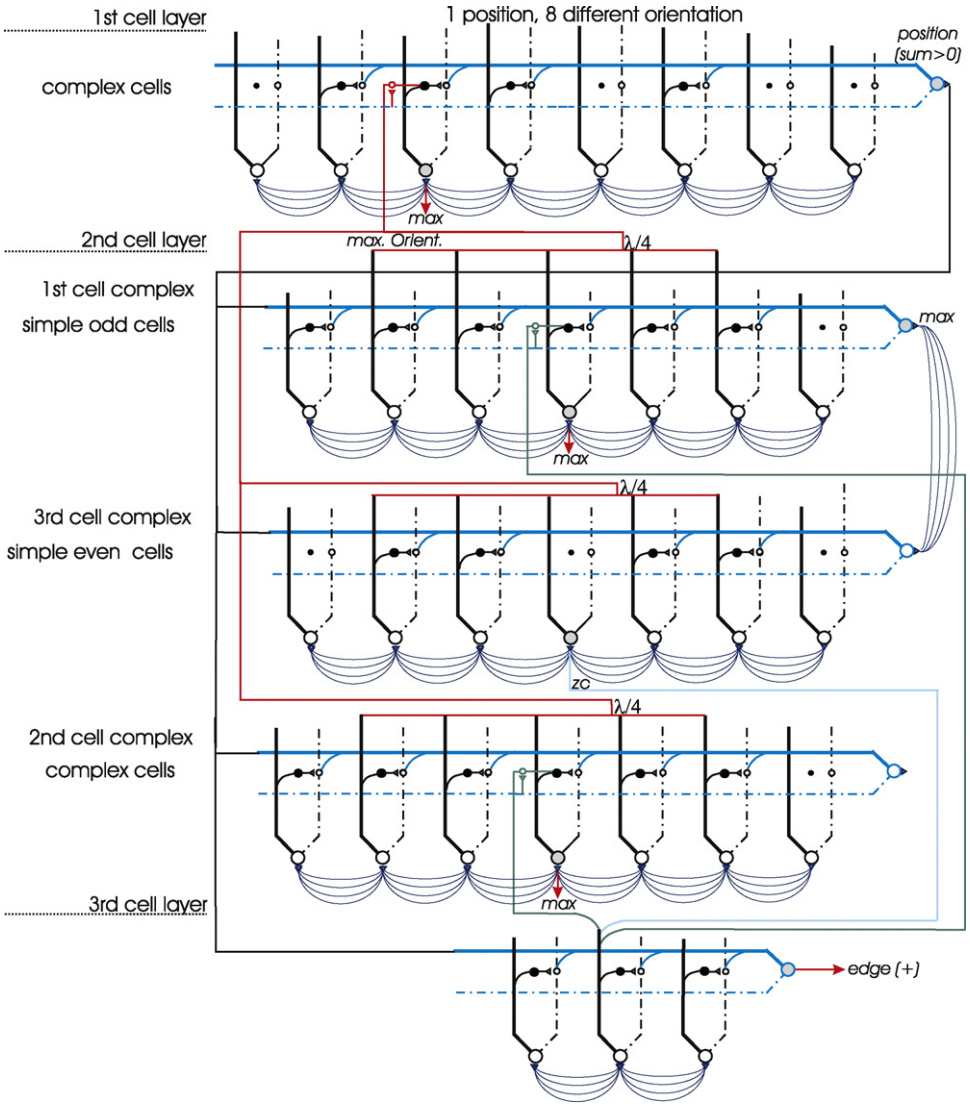

Fig. 2. Schematic diagram for line/edge detection (single event and single scale) using 8 orientations. Cells are represented by solid dots (active cells by big dots), grouping cells by big open circles, and gating cells by small open circles. Dendrites are shown by solid lines and axons by dash-dotted lines (see text).

respectively). Detection accuracy is very good and there remain many small events due to low-contrast textures and the fact that no threshold value has been applied (event amplitudes, for example responses of complex cells at positions where events were detected, are not shown in Fig. 3).

Fig. 4 illustrates continuity processing as applied in the third layer, but in a special context: the detection of good continuity in case of a test image containing even Gabor patches which form a few circular arcs embedded in randomly rotated and slightly shifted (jittered) patches. At each position and for each orientation (only for this image we used $N_\theta = 12$ instead of $N_\theta = 8$), there are many summation cells with different dendritic fields: 12 rotated Gaussian fields in the center and 24 fan-shaped fields around the center. Fig. 4 (top) illustrates the configurations in case of horizontal (x) continuity (simple and complex cells filter in y direction). The center response C is obtained by the summation field $\exp(-(x^2/2\sigma_x^2 + y^2/2\sigma_y^2))$, with $\sigma_x = 5$, $\sigma_y = 3$ and a maximum summation radius $r = 10$ (values in pixels). Fan responses F are obtained by Gaussian weighting from the center, $\exp(-r^2/2\sigma_r^2)$, using $\sigma_r = 10$, a maximum radius of 20, and with angles $\theta_i - \Delta\theta \leq \theta \leq \theta_i + \Delta\theta$, with $\Delta\theta = \pi/N_\theta$ and $\theta_i = i\Delta\theta$. For allowing curvature we consider the center summation and the left and right fans but also the two neighboring fans on each side: to the right these are F_R^{+1} (rotated left), F_R^0 (straight) and F_R^{-1} (rotated right). With 12 orientations and numbering cells tuned to vertical and horizontal orientations 0 and 6, respectively, the corresponding orientations of the complex cells are: 7 in case of F_R^{+1}, 6 in case of F_R^0 and 5 in case of F_R^{-1}. To the left these are F_L^{+1} (rotated left), F_L^0 (straight) and F_L^{-1} (rotated right), which sum responses of complex cells tuned to the same orientations 7, 6 and 5, respectively. Using C in combination with $F_R = \max_i F_R^i$ and $F_L = \max_i F_L^i$ yields the most likely local geometry: completely straight ($F_L^0 - C - F_R^0$), completely curved ($F_L^{-1} - C - F_R^{-1}$ or $F_L^{+1} - C - F_R^{+1}$), combinations of these, and also inflection points (for example $F_L^{-1} - C - F_R^{+1}$). First, the local response is obtained by applying

$$R(x, y) = F_L(x, y) \cdot C(x, y) \cdot F_R(x, y). \tag{14}$$

The processing up to here is done at all orientations i so we have responses $R_i(x, y)$. Second, the global maximum response $R_{\max} = \max_{i,x,y} R_i(x, y)$ is determined for applying a detection threshold at all image positions (a threshold is only applied here in the case of snake continuity). Third, at each position the dominant local orientation is determined by $Id(x, y) = \max_i R_i(x, y)$. Assum-

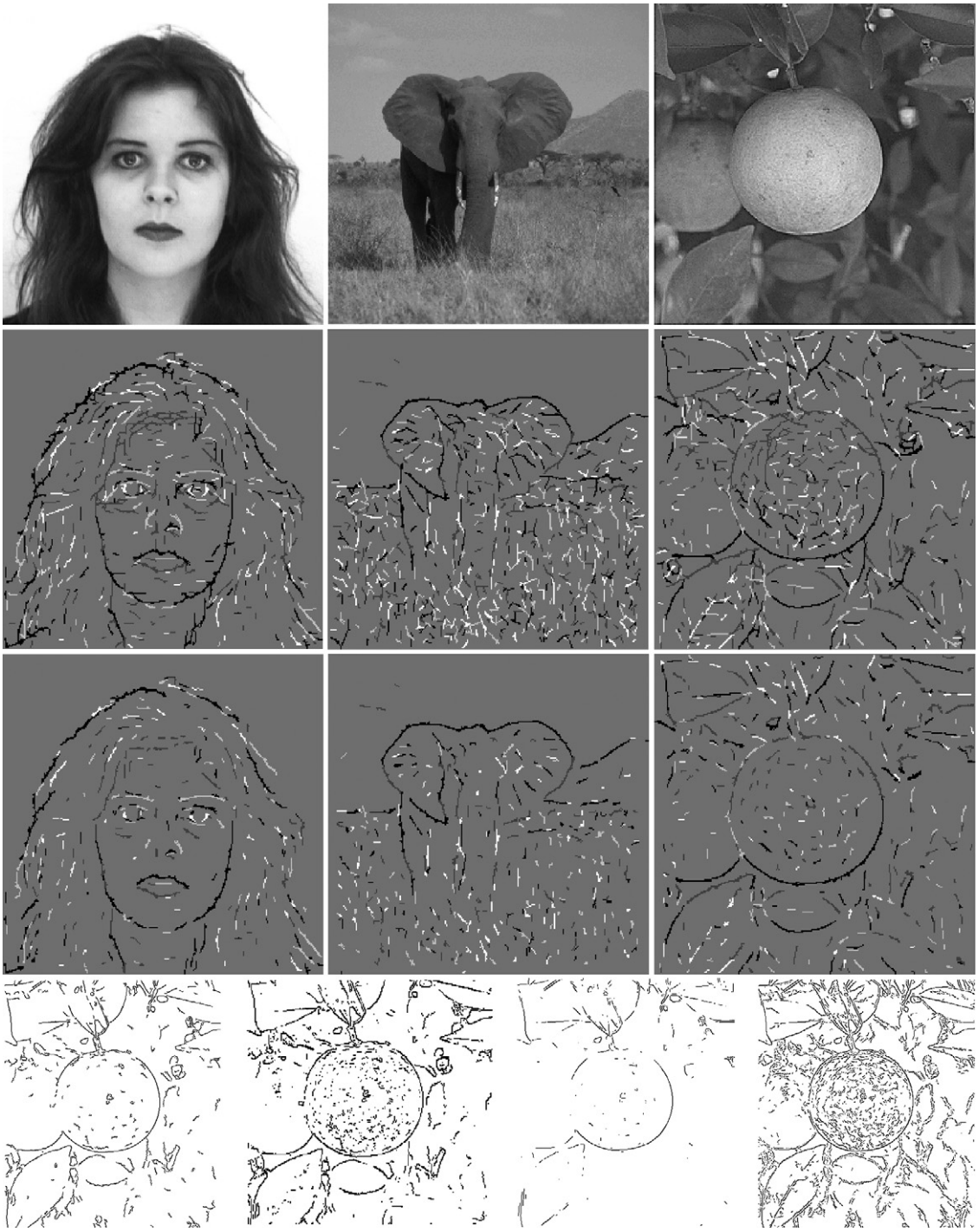

Fig. 3. Line and edge detection at the finest scale. The second row shows positive and negative lines and edges, coded by gray level, without applying NCRF inhibition. The third row shows the same with NCRF inhibition. The bottom row shows edges detected by the (from left to right) Bergholm, Canny, Iverson and Nalwa algorithms in the case of the orange image.

ing $Id(x, y) = 0$ (Fig. 4 top) for simplicity, the final test consists of checking for a local maximum orthogonally (in y) and the global threshold:

$$R(x, y) > 0.1 \cdot R_{\max} \quad \wedge \quad R(x, y) > R(x, y + \delta), \tag{15}$$

with $\delta \in [-\sigma_y, \sigma_y]$ and $\delta \neq 0$. Fig. 4 (bottom-right) shows that most parts of the circular arcs have been detected, but also many more curved and straight lines. We note that the above process is completely data-driven, bottom-up and parallel. Similar processes, including one based on serial processing, are described by Hansen and Neumann (2008) and Roelfsema (2006). We also note that this process, especially at coarse scales, is extremely expensive in terms of CPU time and memory requirements. It depends on the application whether the process can be applied at all scales, with appropriate scaling of the center- and fan-summation parameters, or only at fine scales. Therefore, in this paper continuity processing is applied at all scales but only with very small summation areas.

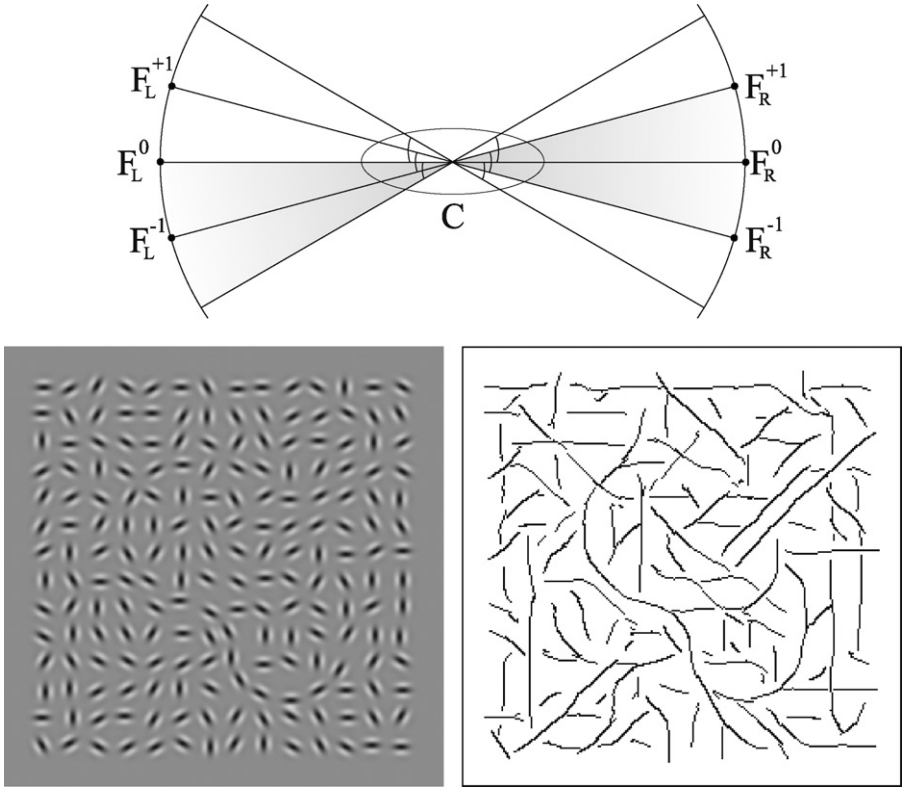

Fig. 4. Top: continuity processing with center- and fan-summation areas, *C* and *F*, in case of horizontal events with curvature (see text). The shading shows combination $F_L^{-1} - C - F_R^0$. Bottom-left: test image with Gabor patches and "snakes" embedded in randomized patches. Bottom-right: detection result.

It has been shown that non-classical receptive field (NCRF) inhibition can be used to suppress information in textured regions (Grigorescu et al., 2003; Rodrigues and du Buf, 2005a). Instead of applying such inhibition only to keypoint detection at fine scales (Rodrigues and du Buf, 2006d), it can also be applied to line and edge detection. The third row in Fig. 3 shows detection results with isotropic (I-NCRF) inhibition applied to responses of complex cells (see Grigorescu et al. (2003); Rodrigues and du Buf (2005a) for the mathematical formulation). As a result, many small events in the face and hair (Fiona), ears and grass (elephant), and orange and tree have been suppressed and the most important events remain. For comparing our results obtained with NCRF inhibition in the case of the elephant image we refer to Grigorescu et al. (2003), but we note that they developed contour (edge) detection algorithms, whereas we can distinguish between edges and lines with different polarities, which is necessary for visual reconstruction; see below. The bottom row in Fig. 3 allows us to compare our results (orange image) with state-of-the-art but edge-only algorithms in computer vision, i.e., Bergholm, Canny, Iverson and Nalwa, see Heath et al. (2000) and also http://marathon.csee.usf.edu/edge/edge_detection.html. This site shows 12 results of each method obtained with different parameter selections. In contrast, our cortical model is applied with no free parameters.

In Sections 6 and 7 we will apply detected lines and edges to object categorization and recognition. It is therefore important to study the stability of detected events as a function of object illumination and background. A change of the background, with arbitrary and complex patterns, will almost always lead to a change of events detected at an object's contour. Positive edges may become negative ones and vice versa. Nevertheless, events will be detected at the same positions – and the use of only events without event type nor polarity will be studied in Section 6 – unless a part of the object at the contour and the neighboring background have the same level of gray or color. We therefore took two images of a plastic cow model with different illuminations, see Fig. 5 (top). The illumination consisted of a very diffuse source in order not to cause a shadow on the background, plus a spotlight which was positioned at azimuth and altitude angles of about 45 degrees, highlighting the rear (Fig. 5 top-left) or the head (top-right). This resulted in different shadows on the floor and low contrast at some contour parts (snout, tail, belly and lower part of the neck). The two images on the middle row show detected events at $\lambda = 4$. As can be seen, there are indeed differences between the two results, at the neck and belly, also the snout, but a big part of the tail contour (right image) has still been detected. Assuming that features of one image are stored in memory, which have to be compared with features of the other image, corresponding features of the two images need to contribute to categorization and recognition. The bottom two images show corresponding features, i.e., in white the same event type and polarity and in black events with different type and polarity. The bottom-left image was generated by using a relaxation area of 3×3 pixels, the bottom-right one with 5×5 pixels. Although parts of snout, neck and belly lack corresponding events, there remain sufficient features – not only events but in most cases also event types and polarities – to conclude that the two objects are the same or at least similar.[1] In conclusion, imaging conditions may change and these may introduce significant differences in images of the same objects, but we can expect that the biggest part of the symbolic object representation in terms of lines and edges, or at least events, will be rather stable. This stability is very important when exploiting the line/edge representation in object categorization and recognition.

[1] Looking at the middle images of Fig. 5, the question arises: cow or bull?

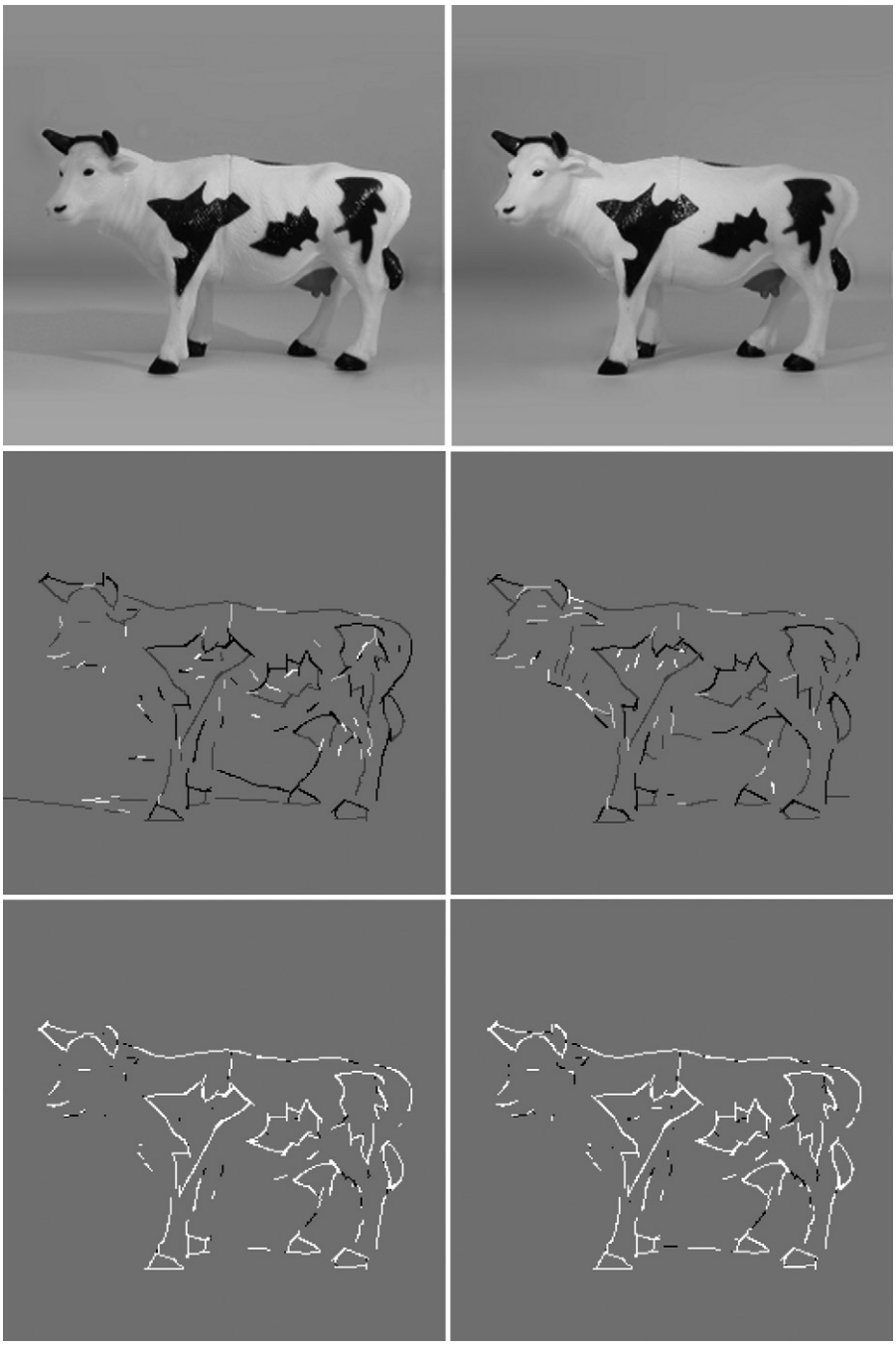

Fig. 5. Top: input images with different spotlights, on the rear (image at left) and on the head (image at right). Middle: detected lines and edges at $\lambda = 4$. Bottom: corresponding features of the middle-row images by applying a relaxation area of 3×3 pixels (left) and 5×5 pixels (right). White shows co-occurrence of equal event type and polarity, black shows co-occurrence of events with different type or polarity.

2.1. Multiple Scales

We now focus on the multi-scale line/edge representation. Although NCRF inhibition can be applied at each scale, we will not do this for two reasons: (a) we want to illustrate line and edge behavior in scale space for applications like categorization, recognition and visual reconstruction, and (b) in many cases a coarser scale, i.e., increased RF size, will automatically eliminate texture detail. For illustrating scale space we can create an almost continuous, linear scaling with hundreds of scales ($\lambda \in [4, 52]$), but here we will present only a few scales in order to show some properties and complications.

The top two rows in Fig. 6 show events detected at five scales in the case of ideal, solid square and star objects. At fine scales (to the left) the edges of the square are detected, as are most parts of the star, but not at the very tips of the star. This illustrates an important difference between normal computer vision and developing cortical models. The latter must be able to construct brightness maps, and at the tips of the star, where two edges converge, there are very fine lines. The same property of detecting a pair of events or only one event as a function of event distance is also exploited in another feature-integration but fine-scale model (see Fig. 5 in Krüger et al., 2007). However, the same effect also occurs at coarser scales because of increasing RF size, until entire

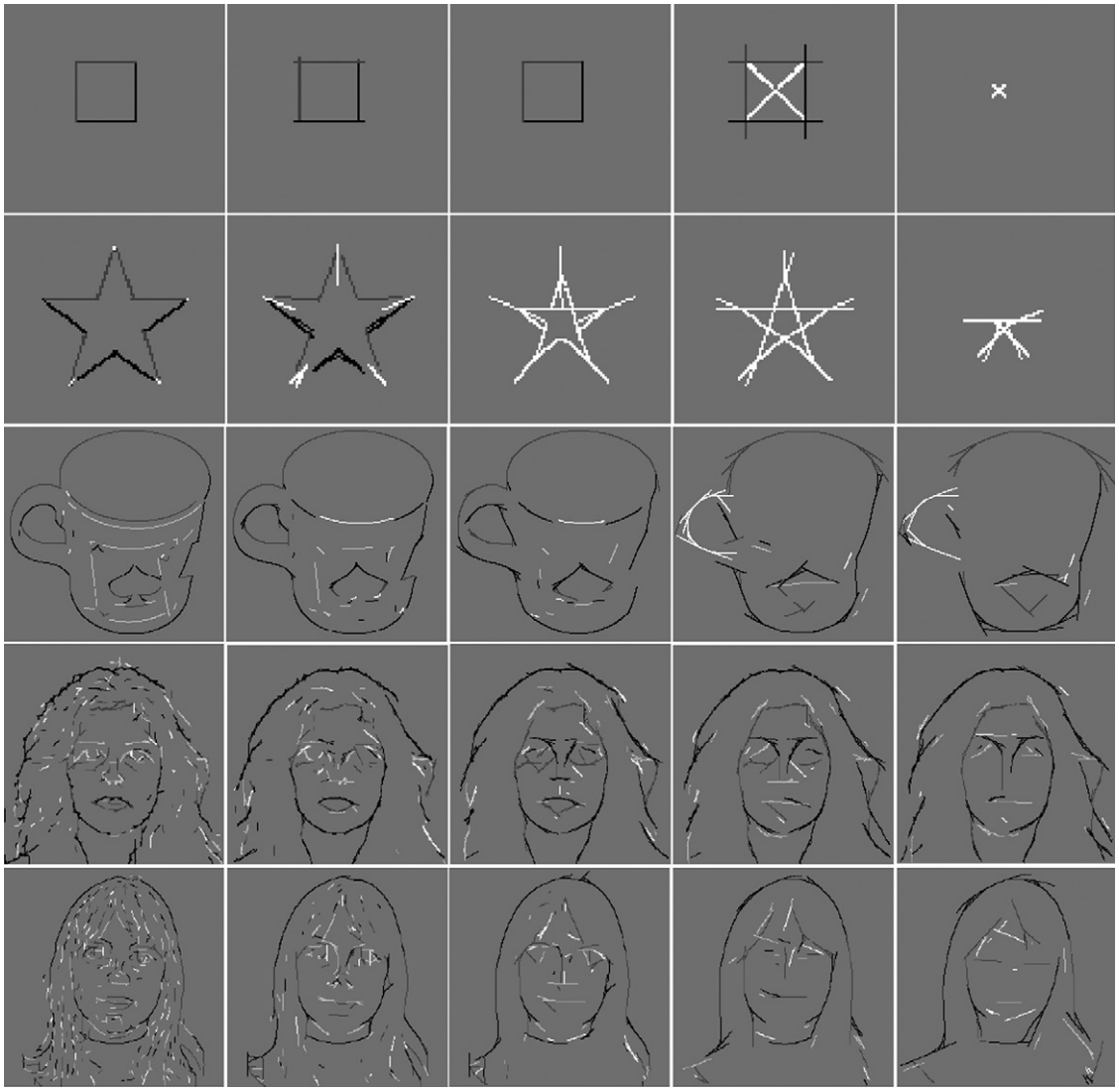

Fig. 6. Top two rows: multi-scale line/edge representations of a square and a star at, from left to right, $\lambda = \{4, 12, 18, 24, 40\}$. Bottom three rows: a mug and two faces at $\lambda = \{4, 8, 12, 24, 28\}$.

triangles are detected as lines and even five pairs of opposite triangles (to the right). In the case of the square, lines will be detected at diagonals, which vanish, with small lengths and amplitudes, at very coarse scales. The third row in Fig. 6 shows a mug, one of the objects that will be used in object categorization, and the bottom two rows show Fiona and Kirsty, two of the images that will be used in face recognition. This figure shows that detail disappears at coarser scales; there the result is more "sketchy" and abstract, a generalization property that will be exploited in object categorization.

Fig. 7 illustrates the concept of stabilization over multiple scales, which will be used in the object recognition model, applying different criteria for scale stability, from top to bottom: single-scale detection without any stability criterium, micro-scale stability over a few neighboring scales (Rodrigues and du Buf, 2004), and stability over 10 and 40 scales with $\Delta\lambda = 5$. This figure shows that many important detected events are rather stable over many scales, which is very important for tasks like visual reconstruction and object recognition.

3. Visual Reconstruction and Brightness Perception

Image reconstruction can be obtained by assuming one lowpass filter plus a complete set of (Gabor) bandpass filters that cover the entire frequency domain, such that an allpass filter is formed—this concept is exploited in wavelet image compression and coding. The goal of our visual system is to detect objects, with no need, nor capacity, to reconstruct a complete image of our visual environment; see change blindness and the limited "bandwidth" of the what and where subsystems (Rensink, 2000). The basic idea is that our physical environment can be seen as external memory (O'Regan, 1992). Yet, the image that we perceive in terms of brightness must somehow be created. A normal image coding scheme, for example by summing responses of simple cells, requires accumulation in one cell layer which contains a brightness map, but this would require "yet another observer" of this map in our brain. In fact, this principle would lead to infinite regress. A simple solution is to assume that detected lines and edges are interpreted symbolically: an active "line cell" is interpreted as having a Gaussian

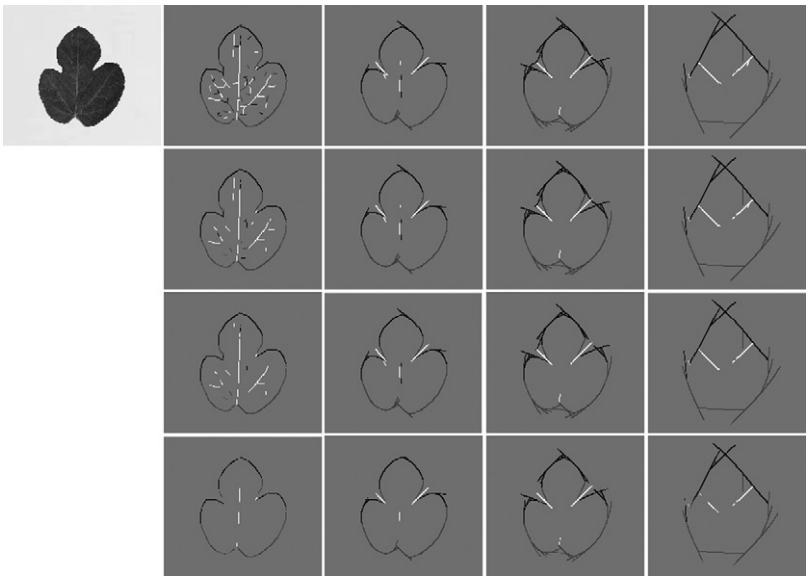

Fig. 7. Left to right: input image (tree leaf) and multi-scale event detection at $\lambda = \{4, 9, 16, 36\}$. Top to bottom: single-scale detection, micro-scale stability, and stability over 10 and 40 scales.

intensity profile with a certain orientation, amplitude and scale, the size of the profile being coupled to the scale of the underlying simple and complex cells. In the case of a vertical line, omitting required rotation to dominant local orientation *Id* for clarity, the profile in x is

$$L_s(x) = \pm\frac{\hat{C}_s(x_0)}{\sigma_s\sqrt{2\pi}}\exp(-(x-x_0)^2/2\sigma_s^2), \tag{16}$$

with $\hat{C}_s$ the amplitude of the complex cell at the line detected at scale s and position x_0, and $\pm$ according to the line's polarity. An active "edge cell" is interpreted the same way, but with a bipolar, Gaussian-truncated, errorfunction profile:

$$E_s(x) = \pm N\frac{\hat{C}_s(x_0)}{\sigma_s\sqrt{2\pi}}\exp(-(x-x_0)^2/2\sigma_s^2)\cdot\mathrm{erf}(x-x_0;\sigma_s), \tag{17}$$

with $N \approx 4.65$ to obtain amplitude normalization (the Gaussian is one in the center, where the errorfunction is zero). As for image coding, this representation must be complemented with a lowpass filter, a process that may exist by means of retinal ganglion cells with big photoreceptive dendritic fields *not* (in)directly connected to rods and cones, the main photoreceptors (Berson, 2003).

One brightness model (du Buf, 1994; du Buf and Fischer, 1995) is based on the symbolic line and edge interpretation. It explains Mach bands (Pessoa, 1996) by the fact that responses of simple cells cannot discriminate lines from ramp edges, and it was shown to be able to predict many brightness illusions such as simultaneous brightness contrast and assimilation, which are two opposite induction effects (the model referred to above was first tested in 1D and has now been extended to 2D).

We only illustrate the symbolic reconstruction process in 2D that will be exploited in face recognition. The left part of Fig. 8 shows, top to bottom, symbolic interpretations of positive and negative edges and lines at fine (left) and coarse (right) scales. The rightmost column illustrates visual reconstruction of the Kirsty image, from top to bottom: input image, lowpass-filtered image (LP_σ), the summation of symbolic line (L_s) and edge (E_s) interpretations (the sum of all images in the left part), and the final reconstruction (R), i.e.,

$$R = \gamma\cdot LP_\sigma + (1-\gamma)\cdot\frac{1}{N_s}\sum_{s=1}^{N_s}(L_s + E_s), \tag{18}$$

with $\gamma = 0.5$. Obviously, the use of more than four scales leads to better reconstructions, but the relative weighting of the lowpass and all the scale components is still under investigation. In principle one can use the same number of scales as used later in the object categorization and recognition processes, for example $N_s = 8$.

There is a big difference between the receptive fields of simple cells, which are always rippled (or waved as in the word wavelet) with possibly multiple wavelengths of the sine and cosine components, and the symbolic interpretations with unipolar Gaussian cross-profiles (lines) and bipolar, Gaussian-truncated errorfunction profiles (edges). In wavelet-based image compression and coding a small error leads to a rippling in the output image, a very disturbing effect for which special postprocessing has been developed in order to reduce it (Ye et al., 2004). In our brain and visual system there are many neurons with rather random looking dendritic and axonal fields and noisy response patterns, and one can question whether, for example, at all retinotopic positions there are simple and complex cells tuned to all necessary orientations and scales. The image that we perceive looks rather stable and complete. Fig. 9 shows what happens when the information is not complete in the case of wavelet coding (at left) and visual reconstruction (at right). Here, wavelet coding is simulated by straightforward summation of responses of simple cells as used in our model, complemented with the same lowpass-filtered image LP_σ:

$$R(x,y) = \gamma\cdot LP_\sigma(x,y) + (1-\gamma)\cdot\sum_{s,i}(R^E_{s,i}(x,y) + R^O_{s,i}(x,y)), \tag{19}$$

again with $\gamma = 0.5$. The top row in Fig. 9 shows reconstructions using all information at six scales and eight orientations. Visual reconstruction (top-right) is already better than wavelet coding (top-left). The bottom row shows reconstructions when 50% of all information is suppressed by random selection. In the case of

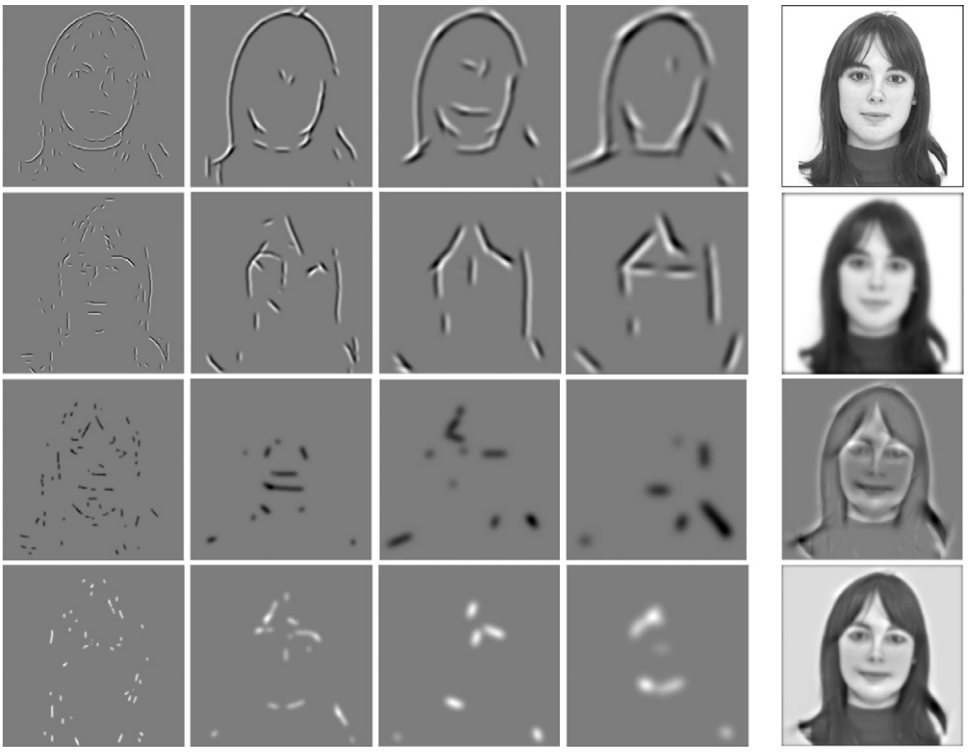

Fig. 8. Left part: multi-scale symbolic line and edge interpretations with, top to bottom, negative and positive edges and lines. Rightmost column: reconstruction of the Kirsty image with, top to bottom: input image, lowpass-filtered image, summation of symbolic line and edge interpretations shown in the left part, and the final reconstruction.

wavelet coding (bottom-left) there does not appear much rippling, which is due to the simple cell model with limited wavelength inside the Gaussian window, but there are disturbing distortions, contrast modulations and a strong "halo" around the mug. In the case of visual reconstruction (bottom-right) one can hardly notice any difference with the complete reconstruction (top-right). There is a very graceful degradation because missing information is provided by neighboring line and edge cells at most scales. An in-depth

Fig. 9. Top: wavelet coding (left) and visual reconstruction (right) using all information (six scales and eight orientations). Bottom: results after randomly suppressing 50% of all information.

analysis of this effect is beyond the scope of this paper and will be reported later, considering extreme suppression rates and different distribution models like Swiss Emmenthal cheese with big holes to study neural degenerations like strabismic amblyopia (Hess et al., 1999; Levi et al., 2007). In any case, Fig. 9 shows that the symbolic retinotopic but distributed line/edge interpretation is a plausible possibility in the visual system whereas the implicit interpretation of distributed simple cells is less likely.

As mentioned above, one of the ideas behind the visual-reconstruction model is the fact that responses of even and odd simple cells do not provide useful information to discriminate lines from ramp edges, such that lines are created at ramp edges but not at ideal, sharp edges, and this explains Mach bands. Indeed, the model tested in the 1D case (du Buf and Fischer, 1995) is one of very few which can predict Mach bands (Pessoa, 1996), but also many other brightness illusions. Among these are simultaneous brightness contrast (SBC) and assimilation (White's effect), two opposite induction effects in which equiluminant (physically identical in luminance) patches are modulated by their neighborhoods: instead of being perceived as having equal brightness, in SBC the neighborhood pushes patch brightness in the opposite direction whereas in assimilation it pulls in the same direction, see Fig. 10 (top). These two effects must somehow be related because of similar neighborhood-patch interactions but a good explanation is still missing. A recent model based on the idea that brightness is largely determined by "the statistical relationship of a particular luminance to all possible luminance values experienced in natural contexts during evolution" (Yang and Purves, 2004) provides an ecological explanation, but if any brightness model is at least required to create Mach bands at ramp edges and no bands at ideal, sharp edges, not to speak of a host of other effects which Yang and Purves (2004) did not consider, this model can be archived together with the numerous models that can predict a few specific effects but not many. Our own model based on visual reconstruction has now been extended with the necessary machinery to process 2D patterns, i.e., the multi-scale line and edge detection and symbolic interpretation, and Fig. 10 (bottom) shows correct model predictions. These are just two examples and many variations, for example the fact that White's effect turns into SBC when bar length is reduced (Moulden and Kingdom, 1989) and many other brightness effects will be reported soon. Important is that our brightness

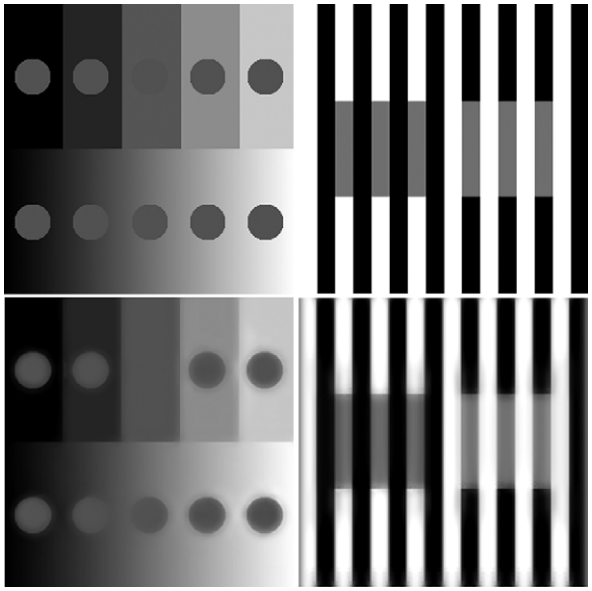

Fig. 10. Top: simultaneous brightness contrast (left) and White's effect (right). All gray circles (left) and bars (right) are equiluminant under homogeneous illumination, but instead of having the same brightness they look different. Bottom: corresponding model predictions.

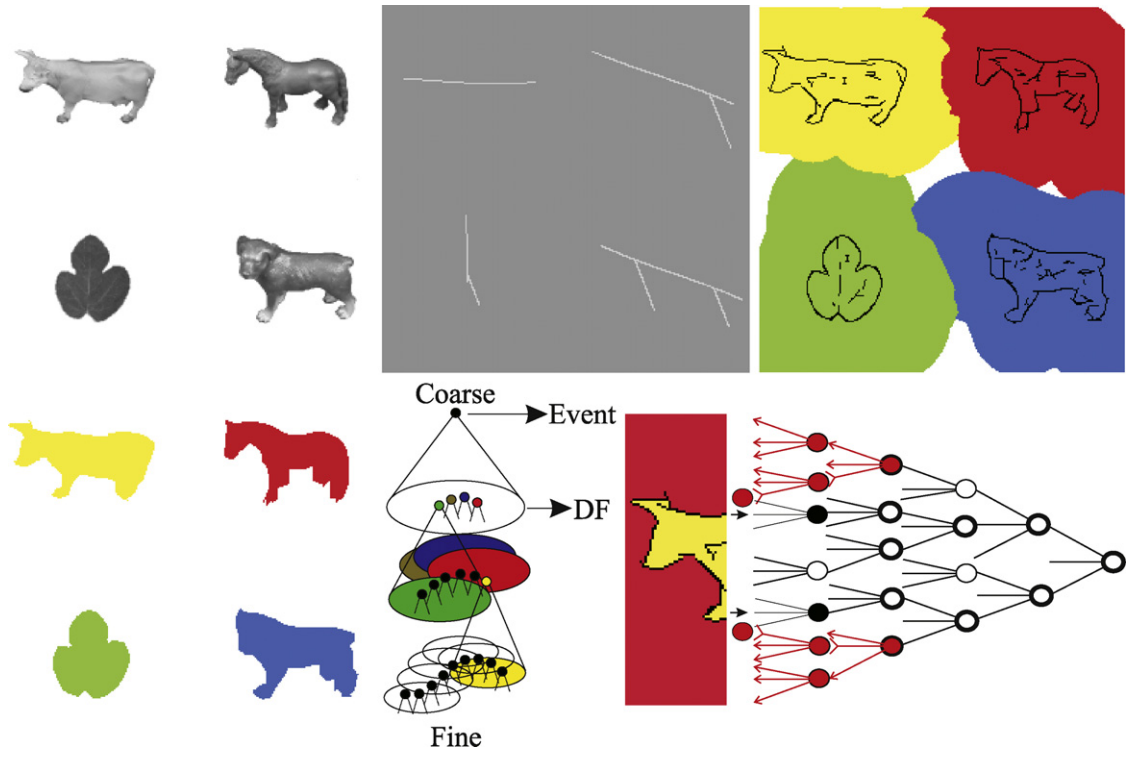

Fig. 11. Object segregation. Top row, left to right: input image with four objects, the representation at $\lambda = 40$, and regions-of-influence with I marking the interior. Bottom row, left to right: result of figure-ground segregation, coarse-to-fine projection (DF denotes dendritic field), and activation and inhibition of grouping cells (right).

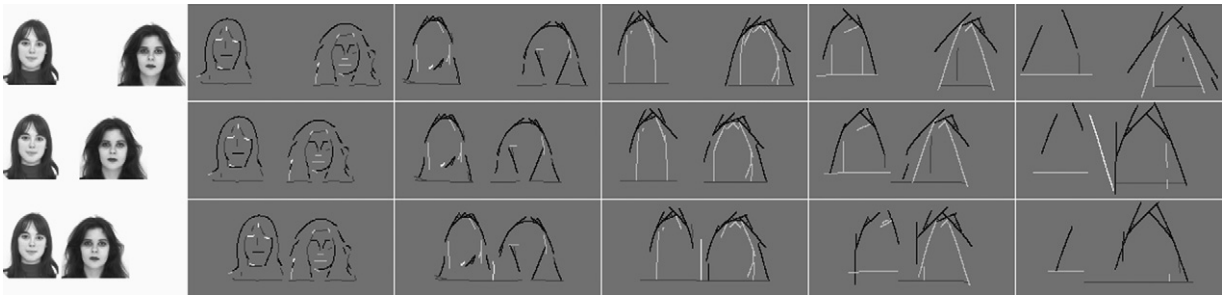

Fig. 12. Object interferences at coarse scales ($\lambda = \{5, 15, 25, 35, 45\}$).

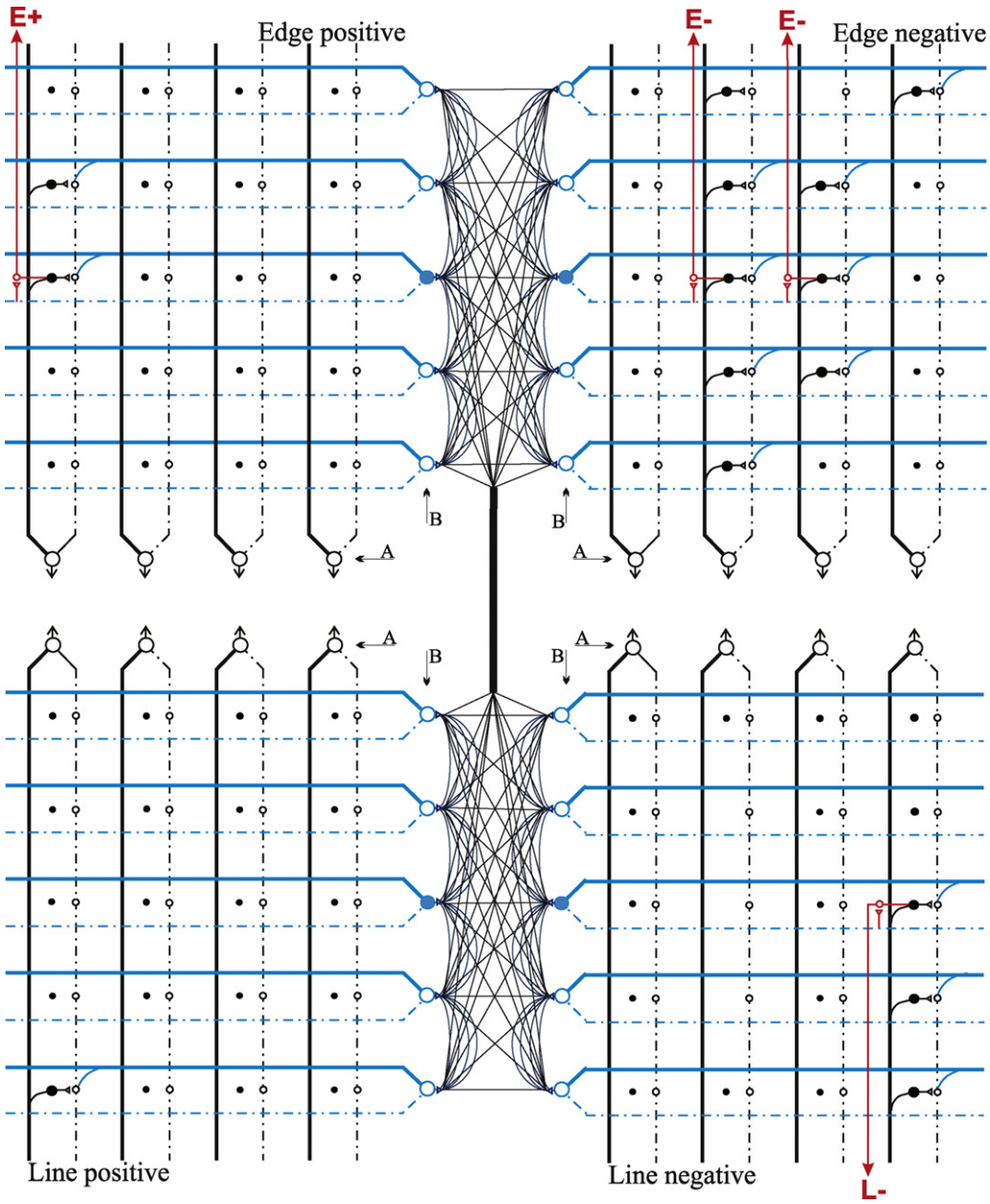

Fig. 13. Schematic diagram for automatic scale selection, with horizontally the position and vertically the scale. Four event maps are used for positive edges (top-left), negative edges (top-right), positive lines (bottom-left) and negative ones (bottom-right). Event cells are represented by solid dots (active event cells by big dots), grouping cells by big open circles, and gating cells by small open circles. Dendrites are shown by solid lines and axons by dash-dotted lines. The positions and scales in the four maps are the same.

model is not based on some scholarly philosophy but on a sound image representation which is also exploited in other tasks: object categorization and recognition.

Summarizing, (1) the multi-scale line and edge interpretation can be used to reconstruct the input image, (2) the symbolic interpretation is rather stable and robust against missing information, and (3) it provides the basis for a simple brightness model which can explain opposite induction effects. The same representation will be used for e.g. face recognition, and the fact that brightness and object recognition may be based on the same image representation has even consequences for explaining consciousness (Section 8).

4. Object Segregation

Until here we have illustrated multi-scale line and edge detection in area V1 and the symbolic interpretation for visual reconstruction in brightness perception, but one of the other goals of the visual cortex is to detect and recognize objects by means of the what and where systems. Object detection and recognition seem like a typical chicken-or-egg problem: without having some idea about the type and characteristics of the object it is not possible to separate the object from its background or from partly occluded other objects. However, a very fast gist system (Rensink, 2000) can be based on feature extractions (color, texture, motion and disparity) and groupings with feed-forward, trained neural networks, and such a system can "bootstrap" the other systems, for example by biasing in parallel possible object templates in associative memory. However, feature-extraction and grouping networks in the gist system may not be able to extract object contours with high precision. We show that high precision can be obtained by linking lines and edges over scales.

Figs. 6 and 7 show typical event maps of different objects, with detail at fine scales and more abstract, "sketchy" information at coarse scales. At a very coarse level, each individual event (group of responding line/edge cells) or connected group of events corresponds to one entire object, see Fig. 11 (top-center). Each event at such a coarse scale is related to events at one finer scale, which can be slightly displaced or rotated, and this continuity continues to fine scales. This relation is modeled by downprojection using grouping cells with a dendritic field (Fig. 11, bottom-center), the size of which defines the region-of-influence. Responding event cells at all scales activate grouping cells, which yields big regions-of-influence (Fig. 11, top-right). This coarse-to-fine-scale process is complemented by inhibition: other grouping cells at the finest scale are activated by responding event cells at that scale and these grouping cells excitate the grouping cells at the one coarser scale but inhibit active grouping cells outwards, as shown in red in Fig. 11(bottom-right). This results in a figure-ground map at the first coarser scale "above" the finest scale (Fig. 11 bottom-left). Results shown were obtained with $\lambda \in [4, 52]$ and $\Delta\lambda = 4$.

A process in V1 as described above can be part of the where system, but it needs to be embedded into a complete architecture with possibly concurrent other processes: object detection, categorization and recognition, such that final segregation may be achieved at the same time as final recognition. The reason is the following: when two objects are very close, they will become connected at coarse scales, see Fig. 12, and separation is only possible by the what system that checks features (lines, edges and keypoints) of individual objects. In other words, object segregation is likely to be driven by "attention" in PF cortex, for example by means of templates that consist of coarse-scale line/edge representations, and this process is related to object categorization.

5. Automatic Scale Selection

Apart from object segregation, other processes may play an important role in the fast where and slower what systems. Concentrating on lines and edges – ignoring other features extracted in V1 – there may be many scales and the tremendous amount of information may not propagate in parallel and at once to IT and PF cortex. It might be useful that lines and edges which are most characteristic for an object are extracted and that these propagate first, for example for a first but coarse object categorization. Such a process may assist or complement surface-based feature extractions in the gist system as discussed in the previous section.

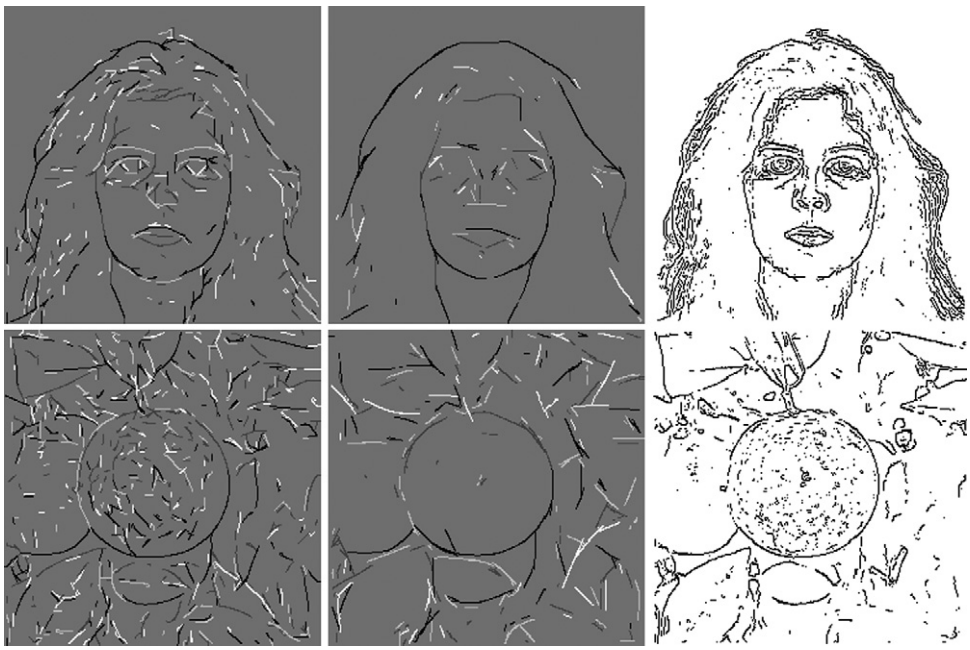

Fig. 14. Automatic scale selection applied to Fiona and orange images. Left: automatic scale selection without a stability criterion. Center: with stability over 20 scales. Right: for comparison results obtained with the Canny edge algorithm.

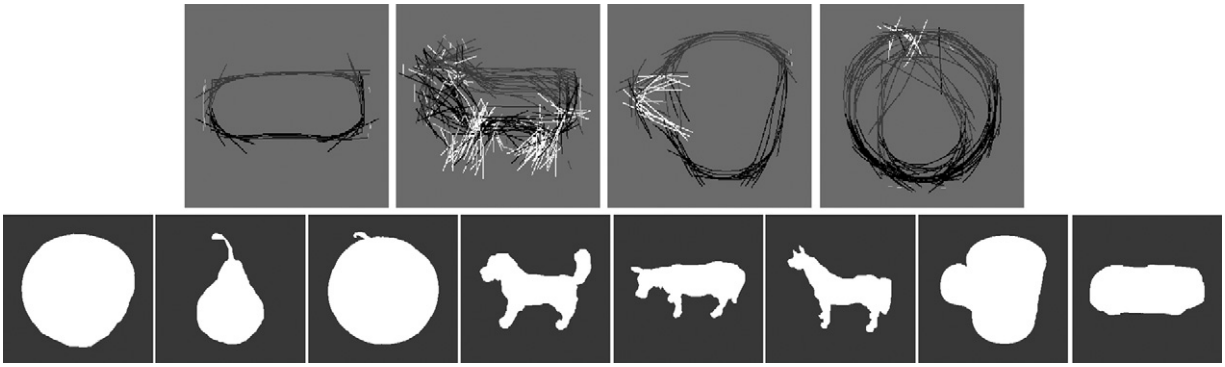

Fig. 15. Top: templates for pre-categorization based on 15 and 5 images at $\lambda = 32$. Bottom: examples of segregated objects.

In Fig. 7 we have seen that different criteria for spatial stability over scales lead to different line/edge selections. Therefore, as was done in the case of keypoints (Rodrigues and du Buf, 2006d), our proposed scheme consists of selecting the scale which counts the maximum number of *stable* events. This can be achieved with a few, simple processes, in which we assume that outputs of event cells are binary.

First, a retinotopic map of grouping cells is assumed. A diagram of event, grouping and gating cells is shown in Fig. 13. This diagram is sub-divided into four parts, with the top-left part for positive edges, the top-right part for negative edges, and similarly the bottom parts for positive and negative lines. In a neural layer the four parts can be mixed if retinotopic mapping is preserved. All four maps show the same positions and scales, with horizontally the position and vertically the scale. The grouping cells marked A have linear dendritic fields (solid black lines) that connect to event cells E (solid dots; active cells are big dots). These grouping cells sum all active event cells at their position, over scale, which yields a sort of histogram: $A(x) = \sum_s E_s(x)$. Second, at each scale, active event cells activate gating cells (triangular synapses next to open circles); these gate the outputs of grouping cells A (black dash-dotted axons) in the "histogram map" at the same position. Third, at each scale, other grouping cells (marked B) sum outputs of all gating cells. The latter grouping cells "count" stable events at all individual scales: $B(s) = \sum_x A(x)E_s(x)$. In words, the B cells count active event cells at each scale, but each event is weighted by the number of active event cells at the event's position. Fourth, the grouping cell with maximum activity is selected (winner takes all: $\hat{s} = \max_s B(s)$) and its axon activates other gating cells that gate outputs of event cells at their scale. The outputs of the latter gating cells provide the map which has the maximum number of stable events $E(x) = E_{\hat{s}}(x)$.

Fig. 14 (at left) shows results of automatic scale selection without an additional stability criterion in the case of the Fiona and orange images. The center images show results when stability over at least 20 scales is required. Many events have disappeared but the

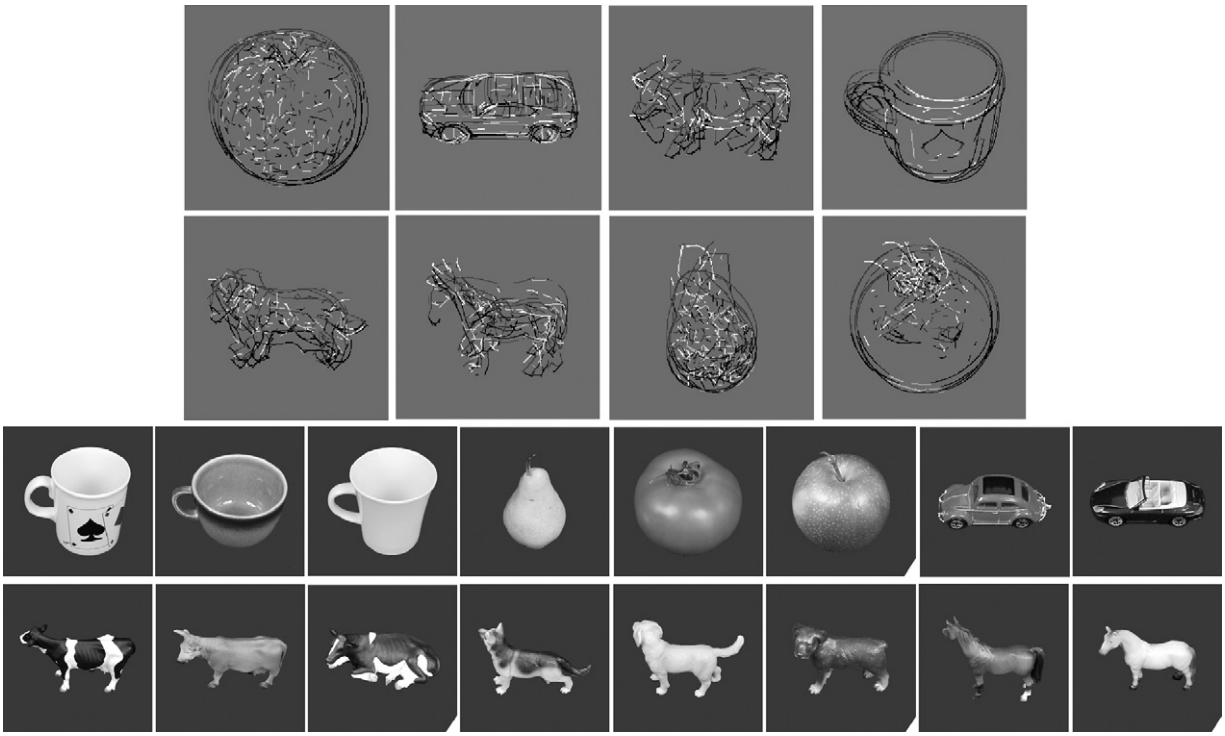

Fig. 16. Top: templates for final categorization based on 5 images at $\lambda = 8$. Bottom: examples of object images, the more difficult ones are marked by a white triangle in the bottom-right corner.

most important ones remain. The right images show, for comparison, some results of Canny's edge detector. Results obtained with other edge detectors can be found in Heath et al. (2000).

6. Object Categorization

Object recognition is a clearly defined task: a certain cat, like the neighbors' red tabby called Toby, is recognized or not. Categorization is more difficult to define because there are different levels, for example (a) an animal, (b) one with four legs, (c) a cat and (d) a red tabby, before deciding between our own red tabby called Tom and his brother Toby living next door. It is as if we were developing categorization by very young children: once they are familiar with the family's cat, every moving object with four legs will be a cat. With age, more features and objects will be added. Here we explain our experiments with a two-level approach: three types of objects (horses, cows, dogs) are first grouped (animal), which we call *pre-categorization*, after which *categorization* determines the type of animal. Instead of creating group templates in memory on the basis of lowpass-filtered images as proposed by the LF model (Oliva et al., 2003; Bar, 2004), we will exploit coarse-scale line and edge templates. In addition, pre-categorization will be based on line and edge templates of contours, i.e. solid objects, available through segregation (Fig. 15), to generalize shape and to eliminate surface detail.

We used the ETH-80 database (Leibe and Schiele, 2003), in which all images are cropped such that they contain only one object, centered in the image, with a 20% border area. Images were rescaled to a size of 256×256 pixels. We selected 10 different images of 8 groups (dogs, horses, cows; apples, pears, tomatoes; cups/mugs and cars), in total 80 images. Fig. 16 shows examples. Because views of objects are also normalized (e.g. all animals with the head to the left), and because different objects within each group are characterized by about the same line/edge representations at coarser scales, group templates can be constructed by combining randomly-selected images. The multi-scale line/edge representation was computed at 8 scales equally spaced on $\lambda \in [4, 32]$.

6.1. Pre-Categorization

Here the goal is to select one of the groups: animal, fruit, cup or car. We used the three coarsest scales with λ equal to 24, 28 and 32 pixels. Group templates were created by combining all images (30 animals, 30 fruits, 10 cups, 10 cars), and by random selections of half (15 and 5) and one third (10 and 3) of all images. By using more images, a better generalization can be obtained, for example the legs of animals can be straight down or more to the front (left). Fig. 15 shows examples of segregated objects and line/edge templates when using half of all images. For each group template, at each of the three scales, a positional relaxation area was created around each responding event cell, by assuming grouping cells with a dendritic field size coupled to the size of underlying complex cells (Bar, 2003). These grouping cells C sum the occurrence of events E in the *input images* I around event positions in the *templates* T, which is a sort of local correlation. Mathematically, at scale s and at all positions $\bar{x}$ where events E of a template T_i are stored in memory, $E_s^{T_i}(\bar{x}) \neq 0$, local grouping cells $C_s^L(\bar{x})$ with circular dendritic fields Δ_s at $\bar{x}$ are activated. These sum events in the input image:

$$C_s^{L,T_i}(\bar{x}) = \sum_{x,y \in \Delta_s} E_s^I(x, y). \tag{20}$$

Then, activities of all activated local grouping cells are grouped together, which yields a sort of global correlation:

$$C_s^{G,T_i} = \sum_{x,y} C_s^{L,T_i}(\bar{x}). \tag{21}$$

Global groupings were summed over scales and the template with the maximum response was selected, i.e., $C^{G,T_i} = \sum_s C_s^{G,T_i}$ and $\hat{T}_i = \max_i C^{G,T_i}$. There may be more scenarios but this one was tested. Without spatial relaxation, most of the model is the same but Δ_s reduces to one (pixel) position and the local grouping is basically the logical AND function.

Table 1 summarizes results (misclassified images) in the form of mean (standard deviation). Obviously, positional relaxation leads to better results when not all images are used in building the templates, and using more images is always better. Using relaxation *and* more images increases shape generalization, however with the risk of running into over-generalization, which did not occur in our tests. On average, different random selections gave very similar results when the three sub-groups (horses/cows/dogs and apples/pears/tomatos) were about equally represented. Most errors occurred, with and without relaxation, between car/animal and cup/fruit. These errors can be explained by the global correlations between the elongated (car/animal) and round (cup/fruit) shapes, see Fig. 15.

6.2. Categorization

After pre-categorization, assuming zero errors, there remains one problem in our test scenario: the animal group must be separated into horse, cow and dog, and the fruit group into apple, pear and tomato. We could have used 6 templates (cups and cars have already been categorized), but we experimented with 8 templates and all 80 images, and applied the multi-scale line/edge maps at all 8 scales (λ equal to 4, 8, 12, 16, 20, 24, 28 and 32) of the real input images (not of the solid, segregated objects). We did this because categorization is supposed to be done *after* pre-categorization, i.e., when also fine-scale information has propagated to IT cortex (see Section 1).

Templates were constructed as above with random selections. Final groupings (global correlations) were compared over the 8 scales and the one with most coherent (maximum) correspondences was selected (in the case of 4–4 we simply took the last one). Table 1 presents results (misclassifications) obtained with positional relaxation.

Again, by using more images in building the templates, generalization is improved and the number of miscategorized images decreases. When using half (5) or even one third (3) of all images, all car and cup images were correctly categorized, and no fruits were categorized as animals and vice versa. Typical miscategorizations were dog/cow, horse/dog, horse/cow and apple/tomato. Fig. 16 shows, apart from examples of images and group templates created by combining 5 images (top), the more difficult images with a white triangle in the bottom-right corner. It should be stressed that this is an extremely difficult test, because no color information

Table 1
Results obtained with pre-categorization and categorization.

	All	Half	Third
Pre-categorization template construction	30/10	15/5	10/3
Error without relaxation	0.0%	5.7%(0.6)	8.0%(1.7)
Error with relaxation	0.0%	3.0%(1.0)	4.3%(0.6)
Categorization template construction	10	5	3
Error with relaxation	0.0%	9.3%(2.1)	12.7%(4.0)

Fig. 17. Examples of images of eleven persons against a dark or bright background with different size normalizations.

has been used and apples and tomatos have the same, round shape. By contrast, all pear images, with a tapered shape, have been correctly categorized. The fact that most problems occurred with the animals was expected, given the small differences of heads, necks and tails (Fig. 16). Categorization is the last step before recognition in which attention shifts to finer scales that reflect minute differences. Nevertheless, only about 9 errors in 80 images (the "50/50 training and testing" scenario) is a very promising starting point for refining the algorithms, for example by using a more hierarchical scenario with more categorization steps, in which attention is systematically steered from the coarsest to the finest scales.

7. Face Recognition

The final goal in vision is object recognition, but here we focus on face recognition by the multi-scale line and edge representations. This completes face detection as presented in the previous paper (Rodrigues and du Buf, 2006d), in which saliency maps and the multi-scale keypoint representation have been used for detecting facial landmarks and thus entire faces. In addition, it was also shown that keypoints can be used for Focus-of-Attention, i.e., to "gate" detected keypoints in associated Regions-of-Interest. The same process can be used to gate detected lines and edges in the Regions-of-Interest. The idea of combining keypoints with lines and edges resembles the bottom-up data streams in the where (FoA) and what (lines/edges) subsystems; for more details see Rodrigues and du Buf (2006b). Of course, this is a simplification because processing is limited to cortical area V1, whereas in reality the two subsystems contain higher-level feature extractions in areas V2, V4, etc. (Hamker, 2005). The same way, top-down data streams are simplified by assuming that stored face templates in memory, that have been built through experience, are limited to lines and edges, and that a few canonical views (frontal, 3/4) are normalized in terms of position, size and rotation: faces are expected to be vertical; for translation, size and rotation invariance see e.g. Deco and Rolls (2004). An additional simplification is the strict attributions of keypoints and lines/edges to the two subsystems: keypoints can also be used in the what system and lines and edges also in the where system.

In our experiments we used 8 primary scales $\lambda_1 = \{4, 8, 12, 16, 20, 24, 28, 32\}$ with $\Delta\lambda_1 = 4$. Each primary scale is supplemented by 8 secondary scales with $\Delta\lambda_2 = 0.125$, such that, for example, $\lambda_{2,\lambda_1=4} = \{4.125, 4.250, \ldots, 5.000\}$. These secondary scales are used for stabilization. The model consists of the following steps:

(A) *Multi-scale line/edge detection and stabilization.* To select the most relevant facial features, detected events must be stable over at least 5 scales in a group of 9 (1 primary plus the 8 secondary scales).

(B) *Construction of four symbolic representation maps.* At each primary scale, stable events (positions) are expanded by Gaussian cross-profiles (lines) and bipolar, Gaussian-truncated error-function profiles (edges), the sizes of which are coupled to the scale of the underlying simple and complex cells; see Fig. 8 (the four leftmost columns). Responses of complex cells are used to determine the amplitudes of the profiles. As a result, each face image is represented by 4 maps at each of the 8 primary scales.

(C) *The recognition process.* We assume that templates (views) of faces are stored in memory and that these have been built through experience. Template images of all persons are randomly selected from all available images: either one frontal view or two views, i.e. one frontal plus one 3/4 view; see also Valentin et al. (1997). Each template in memory is thus represented by 32 line/edge maps (point B above). Two recognition schemes have been tested:

Scheme 1: At each scale, events in the 4 representation maps (the 4 leftmost columns in Fig. 8) of an input image are compared with those in the corresponding maps of a template. Co-occurrences are summed by grouping cells, which yields a sort of event-type and scale-specific correlation, similar to using Eqs. (20) and (21). Then, the outputs of the 4 event-type grouping cells are summed by another grouping cell (correlation over all event types). This results in 8 correlation factors. These factors are compared, scale by scale, over all templates in memory, and the template with the maximum number of co-

Table 2
Results of face recognition, without partial occlusions.

Recogn. scheme	2	2	2	1	1	1	1	Base line
Images	All	Black	White	All	Black	White	Scales	
Frontal view	91.0	90.6	91.5	89.0	86.8	91.5	85.5	71
Frontal plus 3/4	96.0	100.0	91.5	96.0	100.0	91.5	91.8	81

Fig. 18. Occlusion types 1–5 from left to right.

occurrences over the 8 scales will be selected (in the case of equal co-occurrences we simply select the second template).

Scheme 2: Instead of comparing representations scale by scale, only one global co-occurrence is determined by using more levels of grouping cells, i.e., first over maps of specific event types, then over event types, and finally over scales. The template with the maximum is selected by non-maximum suppression.

From the Psychological Image Collection at Stirling University (UK) we selected 100 face images of 26 persons in frontal or frontal-to-3/4 view, with different facial expressions. From those, 13 persons are seen against a dark background, with a total of 53 images, of which 40 images are in frontal view, 11 images are in (very near) 3/4 view (4 persons), and 2 frontal images with added Gaussian and speckle noise (1 person). The other 13 persons (47 images) are seen against a light background, in frontal or near-frontal view. For typical examples see Fig. 17. All persons are represented with at least 3 different facial expressions. In view of the tremendous amount of data already involved in our simple experiments, huge databases cannot (yet) be processed.[2]

All recognition tests involved the entire set of 100 images, although results will also be specified in terms of the subsets of 53 and 47 images in order to analyze the influence of the two different backgrounds and size normalizations. For each person we used two different types of templates: (1) only one frontal view, and (2) two views, frontal and 3/4, but only in the case of 4 persons represented by images in frontal and 3/4 views. In all cases, template images were created by random selection of input images. In order to study robustness with respect to occlusions, a second set of tests was conducted in which partially occluded representations of input images were matched against complete representations of templates.

Table 2 presents the results by testing all images ("all") and by specifying (splitting) these in the case of a dark ("black)" or light ("white") background. The penultimate column "scales" lists the percentage of correct scales that lead to correct recognition in the case of "all" and Scheme 1, where 100% corresponds to 800 because of 8 scales and 100 images. The last column "base line" lists the number of all 100 images that have been recognized with absolute certainty, i.e., when Schemes 1 and 2 and all scales point at the same person.

Comparing columns "all," "black" and "white," there are significant differences because dark and blond hair against dark and light backgrounds cause different events, or even no events, at the outline of the hair. Although the "all" results are reasonably close to the best results, separation of different backgrounds can lead to better but also to worse results. This aspect certainly requires more research, for example with size-normalized templates in memory and dynamic routing of features of unnormalized input faces, such that persons photographed against different backgrounds can be included. Best results were obtained when using two templates with frontal and 3/4 views. Using all events, both recognition schemes yielded a recognition rate of 96%, whereas 81% was the base line with absolute certainty. The difference of 15% is due to relative ranking with some uncertainty. In future research it will make sense to increase the base line, especially when larger databases with more variations will be considered. It should be mentioned that small changes in the hairstyle, or in the face like spectacles (Fig. 17, 3rd row, second from left), or even small pose changes (Fig. 17, 4th row, two leftmost) did not much affect classification, as expected, due to the generalization at coarse scales. However, dramatic changes like the one shown in Fig. 17, 4th row, the five rightmost images, which show Kirsty before and after a change of hairstyle, lead to incorrect results if we consider only one group, but to correct results if we consider two groups, before and after the change.

Our best result of 96.0% is a little bit better than the 94.0% obtained by Petkov et al. (1993a) and the 93.8% by Hotta et al. (2000), and it is very close to the 96.3% reported by Ekenel and Sankur (2005), despite the fact that in all studies the number of tested faces and the databases are different.

In the last experiments we tested the influence of the 5 types of occlusion as shown in Fig. 18, using all 100 images and applying recognition Scheme 2 with templates that combine frontal and 3/4 views. Because of the tremendous amount of storage space (and CPU time) involved, all representations were not re-computed (500 images!) but the occlusions were directly applied to the already computed face representations, thereby suppressing event information in the recognition process. This is an approximation of real occlusions, but it indicates the relative importance of different facial regions in the recognition scheme. Table 3 presents results in terms of "rate (base line)," which must be compared with the bottom part of Table 2, i.e., the first and last columns: 96 (81).

[2] One hundred images of 256×256 pixels, with 72 scales and at each scale 4 representation maps, plus the necessary storage capacity for responses of simple and complex cells necessary for line/edge detection, almost all in four-byte floating-point precision.

Table 3
Results obtained with partial occlusions for the frontal plus 3/4 views

Occlusion type	1	2	3	4	5
Scheme 2	96.0 (80)	95.0 (74)	96.0 (67)	93.0 (64)	97.0 (75)

In the "all events" case and occlusion type 4, instead of 81% only 64% was obtained. But this is the base line: 64 of all 100 images are correctly classified with absolute certainty. The real rate for this occlusion (93%) is very close to the one without occlusion (Table 2, 96%), and slightly worse if compared to the other occlusions. This shows that the multi-scale representation, in particular the shape of the head and hair at the coarser scales, is very robust and contributes most to recognition. The reason for this can be seen in Fig. 8: the stable and "sketchy" information without too much detail at coarse scales. Nevertheless, some contradiction seems to appear when we exclude the eyes (occlusion type 5). In this case we expected a small decrease in performance relative to occlusion types 1–3, but it resulted in the best performance of 97%. An analysis learned that this is due to only one image that failed recognition in occlusion types 1–4 but *not* in type 5. In contrast, the base line is lower, as expected (75 instead of 81). Therefore, the main conclusion is that face and hair contribute about equally to recognition.

8. Consciousness

The fact that object recognition and brightness perception can be based on the same image representation has interesting consequences for consciousness or at least visual awareness. As Crick and Koch (2003) pointed out in their framework, the brain is divided into front and back parts, roughly at the central sulcus, with the front "looking at" the back with most sensory systems, including the visual cortex (in contrast to visual reconstruction in one neural map, as discussed in Section 3, this does not lead to infinite regress). The inherent hierarchical organization is obvious: massively parallel and fast automata for feature extractions operate in the back part, whereas the front part is concerned with slower sensations, attention, free will and the conscious planning of actions. Others, like Denham (2006), try to establish an interpretation at a much lower level: although cognition is a global mental process leading to visual awareness and behavior, feature detection is localized in V1 where dynamic states can be created. Denham illustrates local-to-global processing by local orientation selectivity in the hypercolumns; long-range excitatory connections along contours by the "association fields" of Field et al. (1993) is a higher-level grouping process, and certain structures seen in visual "hallucinations" can be explained by log-polar mapping from retina to cortex. The massive feedback from higher cortical areas suggests a special role of V1: it receives input from LGN into layer 4C. From there are two pathways: (a) to supragranular layers 2-4B and then to extrastriate cortex, where layers 2-4B could be a "blackboard" for conscious visual awareness; and (b) to infragranular layers 5 and 6 and then to other cortical areas, i.e., the non-specific thalamus, subcortical and motor areas, where layers 5–6 could form the basis for a motor-related generative modeling process. This latter explanation of V1 comes close to Koch's local consciousness idea as the first level where perception and action are in a closed loop (Koch, 2008). In addition, visual reconstruction based on multi-scale line/edge interpretations, as explored in Section 3, directly links brightness perception to V1. However, as pointed out above, the front part of the brain is "looking at" the back part, semantic and serial processing takes place there, with attention and at least one higher level of consciousness.

As Rensink (2000) pointed out, the brain has no need nor capacity to analyze all incoming information in order to build a complete, detailed internal representation of our surround: our physical surround *is* our visual memory; see also O'Regan (1992). This idea has even been extended by O'Regan and Noë (2001) to explain the experience of the seeing process as a way of acting, i.e., actively exploring our surround. But they emphasize that what we see is not necessarily what is out there: change blindness is an important clue to attention, awareness and the fact that perhaps as much as 99% of our surround does not provide useful information for the task we are performing.

Let us assume, for the sake of simplicity, only four processing levels: (1) Low-level syntax consisting of multi-scale lines, edges, keypoints plus other features like color and motion. (2) Medium-level complex receptive fields capture entire objects and these are compared in visual memory (a face, who's face?). (3) One level higher, the complex syntax from level 2 (also the simple syntax from level 1?) enters semantic processing. For example, we know that we are in our office, that a colleague is also there and that he or she is looking at us, therefore possibly waiting for a reply. Indeed, cells have been found in the temporal cortex, about halfway between front and back parts and therefore at the edge between global syntactic (level 2) and local semantic (level 3) processing, which respond when someone is looking in our direction, in the case of a specific facial expression, and some cells even prefer direct eye contact (Perrett et al., 1985). (4) At the highest level, semantics concern understanding, overt attention, abstract reasoning and the planning of action.

In this four-level scheme with two syntactic and two semantic levels, visual awareness can be attributed to level 3 and conscious attention etc. to level 4, but where can we put brightness perception? If two versions of the same image of a house are flashed such that no (phi-)motion is evoked, for example one with the chimney to the left and the other with the chimney to the right, in most cases only consciously scrutinizing different regions can reveal the difference. In terms of semantics, the house is a normal one with a chimney, and the position of the chimney adds no useful information, except perhaps for a chimney sweep. Our visual system does what it is supposed to do: it extracts the useful information in terms of meaning, but we may not see what is actually there. Visual reconstruction for brightness perception is based on processing at level 1, which must reflect the two versions of the house and therefore also the difference. But we may perceive only one of the two versions, or the information simply does not "percolate" from the back part of the brain (syntax) to the front part (semantics). Therefore, brightness perception cannot only be a straightforward, data-driven and bottom-up process, because the brightness patterns of the chimney and the sky are different. In other words, as for color (O'Regan and Noë, 2001), brightness may be an illusion which can be manipulated. Obviously, then the same holds for object detection and recognition, although it should be mentioned that change-blindness experiments and demonstrations create a rather artificial context, and that in real life our visual system has no intention to distort reality. All this may simply imply that if we do not pay attention we may not see things and our visual system may assume some solution which could be based on prior experience. The final conclusion, then, is that consciousness is a global process, since it involves local image syntax in V1 which must reflect reality, but also global semantics which may not reflect reality.

In the data flow, from local syntax (the chimney's edges and keypoints in V1), via local geometry (edges and keypoints forming some rectangles) and local semantics (some rectangles on top of a roof; must be a chimney) to global semantics (gist: a normal house), some objects may be lost in transition, because task, attention and context create a gateway and during normal viewing

our brain may not be bothered (blocked!) by processing all image details. Interestingly, the final result of this bottom-up processing, i.e., the gist of the scene, can be linked to Rensink's (2000) triadic architecture in which the gist subsystem is meant to extract, in a fraction of a second, the meaning of a scene and this is used to bootstrap other systems, for example by biasing templates of likely objects stored in memory, like chimneys, for object recognition. Having available models for object segregation, categorization and recognition, including saliency maps for focus-of-attention based on color, texture, also multi-scale keypoints (Itti and Koch, 2001; Parkhurst et al., 2002; Deco and Rolls, 2004; Rodrigues and du Buf, 2006d), as well as models for fast extraction of scene gist (Oliva and Torralba, 2006; Siagian and Itti, 2005), all models can be integrated into a complete architecture and high-level cognitive effects involving task scheduling, context, attention and even consciousness may become subject to explicit modeling.

9. Discussion

Computer vision for realtime applications requires tremendous computational power because all images must be processed from the first to the last pixel. Probing specific objects on the basis of already acquired context may lead to a significant reduction of processing. This idea is based on a few concepts from our visual cortex (Rensink, 2000): (1) our physical surround can be seen as external memory, i.e., there is no need to construct detailed and complete maps, (2) the bandwidth of the what and where systems is limited, i.e., only one object can be probed at any time, and (3) bottom-up, low-level feature extraction is complemented by top-down hypothesis testing, i.e., there is a rapid convergence of activities in dendritic/axonal cell connections from V1 to PF cortex.

In a previous paper we have shown that keypoint scale-space provides very useful information for constructing saliency maps for Focus-of-Attention (FoA), and that faces can be detected by grouping facial landmarks defined by keypoints at eyes, nose and mouth (Rodrigues and du Buf, 2006d). In this paper we have shown that line/edge scale-space provides very useful information for object and face recognition. Obviously, these two representations in V1 complement each other and both can be used in object detection, categorization and recognition. One might think that keypoints provide better information for the fast where system (FoA), whereas lines and edges are better suited for the slower what system. However, both representations are based on responses of simple and complex cells, they may be constructed in parallel, in two different neural layers, and therefore they may be used together. Although there is no psychophysical or neurophysiological evidence for a strict dichotomy, an artificial and dichotomous vision system might be developed, but it must be tested in the context of a complete cortical architecture with ventral and dorsal data streams that link V1 to attention in PF cortex (Deco and Rolls, 2004).

It should be stressed that the fact that face recognition has been explored as a special case of object recognition, using the same symbolic image representation, does not mean that faces and general objects cannot be processed in different ways in the cortex. Indeed, there are several indications that there are fundamental differences (Biederman and Kalocsai, 1997). The facts that faces are very important in our social behavior, and that there are cells which only respond when a face is present in their receptive field, even cells which only respond when there is direct eye contact (Perrett et al., 1985), all point at the existence of a special (and fast) face-processing subsystem, part of which may overlap the machinery for dealing with general objects. The main difference is that faces always have the same geometry, more or less, which was exploited in face detection on the basis of the multi-scale keypoint representation by assuming standard relations between eyes, nose and mouth (Rodrigues and du Buf, 2006d). The same geometry with standard relations between parts does not exist in the case of general (3D) objects. The latter require storage of feature templates in visual memory, covering all possible canonical views. If a view is not available when dealing with an unknown input object, a direct comparison will fail, unless a much more complicated (and conscious) process called mental rotation can provide a solution, or repeated viewing leads to the memorization of the view.

In this paper we presented an improved scheme for line and edge detection in V1, and illustrated the multi-scale representation for visual reconstruction. This representation, in combination with a lowpass filter, yields a reconstruction that is suitable for extending our brightness model (du Buf and Fischer, 1995) from 1D to 2D, for example for modeling brightness illusions.

We also presented a plausible scheme for object segregation, which results in binary, solid objects that can be used to obtain a rapid pre-categorization on the basis of coarse-scale information only. This approach works much better if compared to using lowpass-filtered images, i.e., smeared blobs that lack object-specific characteristics (Oliva et al., 2003; Bar, 2004). Final categorization was tested by using the real objects and more scales, coarse and fine. The results obtained are very promising, taking into account that the tested schemes are extremely simple. Only a fraction of available information, i.e., the line/edge code without amplitude and color information, and without a linking of scales as explored in the segregation model, has been used so far. More extensive tests are being conducted, with more images and objects, concentrating on a linking of scales and a steering of attention from coarse to fine scales. Such improved schemes are expected to yield better results, from very fast detection (where) to slower categorizations (where/what) to recognition (what). The balance between keypoint and line/edge representations in these processes is an important aspect.

The line and edge interpretations at coarser scales lead to stable abstractions of image features (Figs. 6 and 12). This explains, at least partly, the generalization which is required to classify faces with noise, spectacles, and relatively normal expressions and views (Fig. 17). It should be stressed that the recognition scheme is not yet complete, because a hierarchical linking from coarse to fine scales, as already applied in the detection/segregation process, has not been applied. Such an extension can lead to better recognition rates, especially when multiple views (frontal, 3/4 and lateral) of all persons are included as templates in memory. In addition, the multi-scale keypoint representation, which has been ignored here, will contribute very important information.

A new disparity model (Rodrigues and du Buf, 2004) is based on line/edge detection in combination with the linear responses of odd simple cells (Gabor kernel with sine phase) around the center of the receptive fields. Although still at an initial development stage because it must be extended to multi-scale processing, it will be able to directly attribute depth to lines and edges, thereby creating a 3D "wireframe" representation. Such a wireframe representation is used in modeling solid objects in computer graphics. The fact that projections from left and right eyes are very close in the cortical hypercolumns and that many simple and complex cells are also disparity tuned suggests that our visual system processes 3D objects in the same way, probably simplifying 3D object recognition. However, how this is achieved is still an open question. For example, if a vertical edge is detected in left and right images by means of simple and complex cells tuned to horizontal disparity, and depth is directly attributed to it, depth is a local property along the edge if the surface of the object – and therefore also the edge – is curved in depth. Edges which are not vertical are detected by

cells tuned to other orientations, and only the disparity component orthogonal to the edge can be attributed to them (similar to the aperture problem in motion detection). Here we neglect residual responses of "vertical" cells to non-vertical edges and (small) contributions from vertical disparity (Read and Cumming, 2006), i.e., no depth at all can be attributed to horizontal edges. Furthermore, the correspondence problem must be solved (which edge in the left image corresponds to which one in the right image), and receptive field sizes must be taken into account (cells with small receptive fields can only handle small differences in disparity). Both problems can be solved by hierarchical processing, going from coarse to fine scales. As shown by Krüger et al. (2007), lines/edges and keypoints, being 1D and 2D singularities, can be combined into meaningful object primitives, even into 3D primitives in case of stereo vision, yielding very powerful descriptors at high-level syntactic (or low-level semantic) level. This facilitates a disambiguation process in case of moving objects and provides direct input for high-level cognitive tasks, for example a robot manipulating 3D objects. Returning to disparity, by definition keypoints are detected at vertices where lines and edges cross or connect. This holds for all scales, even for coarse ones at which image content is distorted because of large receptive field sizes and response interference effects (du Buf, 1993). Nevertheless, as shown in (Rodrigues and du Buf, 2006d), an analysis of the local neighborhood around detected keypoints provides information to determine the keypoint type, for example L, T or + junction, even the 1D event types at the keypoint. Such annotated keypoints, in line with the approach by Krüger et al. (2007), provide much richer information for linking left and right images than mere clouds of keypoints. Therefore, a possible solution to all problems is to match annotated keypoints in left and right images, going from coarse to fine scales, to extract disparity from corresponding keypoints, and to interpolate depth along lines and edges. Or, more precise, to complement depth information as extracted directly from lines and edges, not forgetting that depth from keypoints at coarse scales may not have corresponding keypoints, nor lines and edges, at small scales. In other words, there may be a sort of depth-diffusion process which starts spanning a 3D surface at a coarse scale which is progressively refined toward finer scales. This proposed approach may provide a precise and stable solution in computer vision, also in case of motion detection, although there are no signs (yet) of similar processes in the visual cortex.

All multi-scale processing and the representations, including keypoints, are restricted to areas V1 and V2. On the other hand, the Deco and Rolls (2004) scheme with ventral and dorsal data streams, necessary for obtaining position and size invariance through projections via areas V2, V4 etc., is solely based on responses of simple cells. In the future, this scheme must be based on features extracted in V1, and further multi-scale processing can be added in the higher areas V2 to PF cortex. We expect that such extensions in adaptive up and down projections will lead to much better results.

Acknowledgments

The authors thank the anonymous reviewers for their comments which helped to improve the manuscript. This research is partly supported by the Foundation for Science and Technology FCT (ISR/IST pluri-annual funding) through the POS-Conhecimento Program which includes FEDER funds, and by the FCT project PTDC/EIA/73633/2006 – SmartVision: active vision for the blind. The orange image is available at http://marathon.csee.usf/edge/edge_detection.html; the elephant image is available at http://www.cs.rug.nl/∼imaging/databases/contour_database/contour_2.html; the ETH-80 database (objects) is available athttp://www.mis.informatik.tu-darmstadt.de/Research/Projects/categorization/download.html; and the Stirling face images are available at http://pics.psych.stir.ac.uk/.

References

Ban, S., Skin, J., Lee, M., 2003. Face detection using biologically motivated saliency map model. Proc. Int. Joint Conf. Neural Netw. 1, 119–124.
Barth, E., Zetzsche, C., Krieger, G., 1998. Endstopped operators based on iterated nonlinear center-surround inhibition. Human Vision Electronic Imaging, SPIE 3299, 67–78.
Bar, M., 2003. A cortical mechanism for triggering top-down facilitation in visual object recognition. J. Cogn. Neurosci. 15 (4), 600–609.
Bar, M., 2004. Visual objects in context. Nature Rev.: Neurosci. 5, 619–629.
Berson, D., 2003. Strange vision: ganglion cells as circadian photoreceptors. Trends Neurosci. 26 (6), 314–320.
Biederman, I., Kalocsai, P., 1997. Neurocomputational bases of object and face recognition. Philosoph. Trans. R. Soc.: Biol. Sci. 352, 1203–1219.
Bruce, V., Green, P., Georgeson, M., 2000. Visual perception. In: Physiology, Psychology and Ecology. Psychology Press Ltd, UK.
Crick, F., Koch, C., 2003. A framework for consciousness. Nature Neurosci. 6, 119–126.
Csurka, G., Bray, C., Dance, C., Fan, L., 2004. Visual categorization with bags of keypoints. In: Proc. Int. Worksh. Statistical Learning in Comp. Vision, Prague (Czech Republic), pp. 1–16.
Deco, G., Rolls, E., 2004. A neurodynamical cortical model of visual attention and invariant object recognition. Vision Res. 44 (6), 621–642.
Delorme, A., Thorpe, S., 2001. Face identification using one spike per neuron: resistance to image degradations. Neural Netw. 14 (6–7), 795–804.
Denham, M., 2006. The role of the neocortical laminar microcircuitry in perception, cognition, and consciousness. In: Talk in AISB'06 GC5 Symposium.
du Buf, J., Fischer, S., 1995. Modeling brightness perception and syntactical image coding. Optical Eng. 34 (7), 1900–1911.
du Buf, J., 1993. Responses of simple cells: events, interferences, and ambiguities. Biol. Cybern. 68, 321–333.
du Buf, J., 1994. Ramp edges, Mach bands, and the functional significance of simple cell assembly. Biol. Cybern. 70, 449–461.
du Buf, J., 2001. Modeling brightness perception. In: van den Branden Lambrecht, C.J. (Ed.), Vision models and applications to image and video processing. Kluwer Academic, pp. 21–36 (Chapter 2).
du Buf, J., 2007. Improved grating and bar cell models in cortical area V1 and texture coding. Image Vision Comput. 25 (6), 873–882.
Ekenel, H., Sankur, B., 2005. Multiresolution face recognition. Image Vision Comput. 23 (5), 469–477.
Elder, J., Zucker, S., 1998. Local scale control for edge detection and blur estimation. IEEE Tr. PAMI 20, 699–716.
Field, D., Hayes, A., Hess, R., 1993. Contour integration by the human visual system: evidence for a local "association field". Vision Res. 33 (2), 173–193.
Fleet, D., Jepson, A., Jenkin, M., 1991. Phase-based disparity measurement. Image Understand. 53 (2), 198–210.
Grigorescu, C., Petkov, N., Westenberg, M., 2003. Contour detection based on nonclassical receptive field inhibition. IEEE Tr. IP 12 (7), 729–739.
Hamker, F., 2005. The reentry hypothesis: the putative interaction of the frontal eye field, ventrolateral prefrontal cortex, and areas V4, IT for attention and eye movement. Cerebral Cortex 15, 431–447.
Hansen, T., Neumann, H., 2008. A recurrent model of contour integration in primary visual cortex. J. Vision 8 (8), 1–25.
Heath, M., Sarkar, S., Sanocki, T., Bowyer, K., 2000. A robust visual method for assessing the relative performance of edge-detection algorithms. IEEE Tr. PAMI 19 (12), 1338–1359.
Heitger, F., Rosenthaler, L., von der Heydt, R., Peterhans, E., Kubler, O., 1992. Simulation of neural contour mechanisms: from simple to end-stopped cells. Vision Res. 32 (5), 963–981.
Heitger, F., von der Heydt, R., 1993. A computational model of neural contour processing: figure-ground segregation and illusory contours. In: Proc. Int. Conf. Comp. Vision, Berlin (Germany), pp. 32–40.
Hess, R., Wang, Y., Demanins, R., Wilkinson, F., Wilson, H., 1999. A deficit in strabismic amblyopia for global shape detection. Vision Res. 39, 901–914.
Hotta, K., Mishima, T., Kurita, T., Umeyama, S., 2000. Face matching through information theoretical attention points and its applications to face detection and classification. In: IEEE Proc. 4th Int. Conf. Automatic Face and Gesture Recogn, pp. 34–39.
Hubel, D., 1995. Eye, Brain and Vision, Scientific American Library.
Hupe, J., James, A., Girard, P., Lomber, S., Payne, B., Bullier, J., 2001. Feedback connections act on the early part of the responses in monkey visual cortex. J. Neurophysiol. 85 (1), 134–144.
Itti, L., Koch, C., 2001. Computational modeling of visual attention. Nature Rev.: Neurosci. 2 (3), 194–203.
Koch, C., 2008. The Neuroscience of Consciousness. In: Fundamental Neuroscience, third ed. Elsevier, pp. 1223–1237 (Chapter 53).
Kovesi, P., 1999. Image features from phase congruency. J. Comp. Vision Res. 1 (3), 2–27.
Krüger, N., Peters, G., 2000. Orassyll: Object recognition with autonomously learned and sparse symbolic representations based on metrically organized local line detectors. Comp. Vision Image Understand. 77, 49–77.

Krüger, N., Pugeault, N., Wörgötter, F., 2007. Multi-modal primitives: Local, condensed, and semantically rich visual descriptors and the formalisation of contextual information. IEEE Trans PAMI, submitted for publication (also available as Technical Report no. 2007–4, Robotics Group, The Maersk Mc-Kinney Moller Institute, Univ. of Southern Denmark).

Kruizinga, P., Petkov, N., 1995. Person identification based on multiscale matching of cortical images. Proc. Int. Conf. and Exhib. High-Perf. Comp. Netw. Springer LNCS 919, 420–427.

Lee, T., 1996. Image representation using 2D Gabor wavelets. IEEE Tr. PAMI 18 (10), 959–971.

Leibe, B., Schiele, B., 2003. Analyzing appearance and contour based methods for object categorization. IEEE Proc. Int. Conf. Comp. Vision Patt. Recogn. 2, 409–415.

Levi, D., Yu, C., Kuai, S., Rislove, E., 2007. Global contour processing in amblyopia. Vision Res. 47 (4), 512–524.

Lindeberg, T., 1994. Scale-Space Theory in Computer Vision. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Miikkulainen, R., Bednar, J., Choe, Y., Sirosh, J., 2005. Computational maps in the visual cortex. Springer Science Business+Media, Inc.

Moulden, B., Kingdom, F., 1989. White's effect: a dual mechanism. Vision Res. 29 (9), 1245–1259.

Ohzawa, I., DeAngelis, G., Freeman, R., 1997. Encoding of binocular disparity by complex cells in the cat's visual cortex. J. Neurophysiol. 18 (77), 2879–2909.

Oliva, A., Torralba, A., Casthelano, M., Henderson, J., 2003. Top-down control of visual attention in object detection. IEEE Proc. Int. Conf. Image Process. 1, 253–256.

Oliva, A., Torralba, A., 2006. Building the gist of a scene: the role of global image features in recognition. Progress Brain Res.: Visual Percept. 155, 23–26.

O'Regan, J., Noë, A., 2001. A sensorimotor account of vision and visual consciousness. Behav. Brain Sci. 24 (5), 939–1011.

O'Regan, J., 1992. Solving the "real" mysteries of visual perception: the world as an outside memory. Can. J. Psychol. 46 (3), 461–488.

Parkhurst, D., Law, K., Niebur, E., 2002. Modelling the role of salience in the allocation of overt visual attention. Vision Res. 42 (1), 107–123.

Perrett, D., Smith, P., Potter, D., Mistlin, A., Head, A., Milner, D., Jeeves, M., 1985. Visual cells in the temporal cortex sensitive to face view and gaze direction. Proc. R. Soc. Lond. Series B, Biol. Sci. 223 (1232), 293–317.

Pessoa, L., 1996. Mach bands: how many models are possible? Recent experimental findings and modeling attemps. Vision Res. 36, 3205–3227.

Petkov, N., Kruizinga, P., Lourens, T., 1993a. Biologically motivated approach to face recognition. In: Proc. Int. Worksh. Artif. Neural Netw, pp. 68–77.

Petkov, N., Kruizinga, P., 1997. Computational models of visual neurons specialised in detection of periodic and aperiodic visual stimuli. Biol. Cybern. 76, 83–96.

Petkov, N., Lourens, T., Kruizinga, P., 1993b. Lateral inhibition in cortical filters. In: Proc. Int. Conf. Dig. Signal Proc. and Int. Conf. on Comp. Appl. to Eng. Sys, Nicosia (Cyprus), pp. 122–129.

Qian, N., 1997. Binocular disparity and the perception of depth. Neuron 18, 359–368.

Rasche, C., 2005. The making of a neuromorphic visual system. Springer.

Read, J., Cumming, B., 2006. Does depth perception require vertical-disparity detectors? J. Vision 6, 1323–1355.

Rensink, R., 2000. The dynamic representation of scenes. Visual Cogn. 7 (1–3), 17–42.

Riesenhuber, M., Poggio, T., 2000. CBF: A new framework for object categorization in cortex. In: IEEE Proc. Int. Worksh. Biol. Motivated Comp. Vision, Seoul (Korea), May 15–17, pp. 1–9.

Rodrigues, J., du Buf, J., 2004. Visual cortex frontend: integrating lines, edges, keypoints and disparity. Proc. Int. Conf. Image Anal. Recogn., Porto (Portugal), Springer LNCS 3211, pp. 664–671.

Rodrigues, J., du Buf, J., 2005a. Improved line/edge detection and visual reconstruction. In: Proc. 13th Portuguese Comp. Graphics Meeting, Vila Real (Portugal), pp. 179–184.

Rodrigues, J., du Buf, J., 2005b. Multi-scale keypoints in V1 and face detection. Proc. 1st Int. Symp. Brain, Vision and Artif. Intell., Naples (Italy), Springer LNCS 3704, pp. 205–214.

Rodrigues, J., du Buf, J., 2006a. Cortical object segregation and categorization by multi-scale line and edge coding. Proc. Int. Conf. Comp. Vision Theory Applicat., Setúbal (Portugal) 2, 5–12.

Rodrigues, J., du Buf, J., 2006b. Face recognition by cortical multi-scale line and edge representations. Proc. Int. Conf. Image Anal. Recogn., Póvoa do Varzim (Portugal), Springer LNCS 4142, pp. 329–340.

Rodrigues, J., du Buf, J., 2006c. Face segregation and recognition by cortical multi-scale line and edge coding. In: Proc. 6th Int. Worksh. Patt. Recogn. in Information Systems, Paphos (Cyprus), May 23–24, pp. 5–14.

Rodrigues, J., du Buf, J., 2006d. Multi-scale keypoints in V1 and beyond: object segregation, scale selection, saliency maps and face detection. BioSystems 86, 75–90, doi:10.1016/j. biosystems.2006.02.019.

Roelfsema, P., 2006. Cortical algorithms for perceptual grouping. Ann. Rev. Neurosci. 29, 203–227.

Siagian, C., Itti, L., 2005. Gist: a mobile robotics application of context-based vision in outdoor. In: Proc. IEEE-CVPR Worksh. Attention and Performance in Computer Vision, San Diego, California, pp. 1–7.

Smith, S., Brady, J., 1997. Susan - A new approach to low level image processing. Int. J. Comp. Vision 23 (1), 45–78.

Valentin, D., Abdi, H., Edelman, B., 1997. What represents a face? A computational approach for the integration of physiological and psychological data. Perception 26 (10), 1271–1288.

van Deemter, J., du Buf, J., 2000. Simultaneous detection of lines and edges using compound Gabor filters. Int. J. Patt. Recogn. Artif. Intell. 14 (6), 757–777.

Verbeek, P., van Vliet, L., 1992. Line and edge detection by symmetry filters. In: Proc. 11th Int. Conf. Patt. Recogn. III, pp. 749–753.

Yang, M., Kriegman, D., Ahuja, N., 2002. Detecting faces in images: a survey. IEEE Tr. PAMI 24 (1), 34–58.

Yang, Z., Purves, D., 2004. The statistical structure of natural light patterns determines perceived light intensity. Proc. Natl. Acad. Sci. U.S.A. 101, 8745–8750.

Ye, S., Sun, Q., Chang, E., 2004. Edge directed filter based error concealment for wavelet-based images. IEEE Proc. Int. Conf. Image Process. 2, 809–812.

Zhaoping, L., 2003. V1 mechanisms and some figure-ground and border effects. J. Physiol., 503–515.

Zhao, W., Chellappa, R., Rosenfeld, A., Phillips, P., 2003. Face recognition: a literature survey. ACM Comput. Surveys 35 (4), 399–458.