Multi-scale keypoints in V1 and beyond: object segregation, scale selection, saliency maps and face detection

João Rodrigues [a], J.M.Hans du Buf [b]

[a] *University of Algarve, Escola Superior de Tecnologia, Campus da Penha, 8000-810 Faro, Portugal*

[b] *University of Algarve, Vision Laboratory (FCT), Campus de Gambelas, 8000-810 Faro, Portugal*

Abstract

End-stopped cells in cortical area V1, which combine outputs of complex cells tuned to different orientations, serve to detect line and edge crossings, singularities and points with large curvature. These cells can be used to construct retinotopic keypoint maps at different spatial scales (Level-of-Detail). The importance of the multi-scale keypoint representation is studied in this paper. It is shown that this representation provides very important information for object recognition and face detection. Different grouping operators can be used for object segregation and automatic scale selection. Saliency maps for Focus-of-Attention can be constructed. Such maps can be employed for face detection by grouping facial landmarks at eyes, nose and mouth. Although a face detector can be based on processing within area V1, it is argued that such an operator must be embedded into dorsal and ventral data streams, to and from higher cortical areas, for obtaining translation-, rotation- and scale-invariant detection.

Key words: visual cortex, keypoints, object segregation, automatic scale selection, saliency, Focus-of-Attention, face detection

1 Introduction

Our visual system can still be seen as a huge puzzle with a lot of missing pieces. Even in the first processing layers in area V1 of the visual cortex there

Email addresses: jrodrig@ualg.pt (João Rodrigues), dubuf@ualg.pt (J.M.Hans du Buf).

remain many gaps, despite all knowledge already compiled (Rasche, 2005; Hubel, 1995; Bruce et al., 2000). Nevertheless, some of the gaps are being filled by developing and studying computational models. Models of simple, complex and end-stopped cells have been developed more than ten years ago (Heitger et al., 1992). Several inhibition models (Grigorescu et al., 2003; Petkov et al., 1993), keypoint detection (Heitger et al., 1992; Rodrigues and du Buf, 2004b; Würtz and Lourens, 2000) and line/edge detection schemes (Grigorescu et al., 2003; Rodrigues and du Buf, 2004b; Elder and Sachs, 2004; van Deemter and du Buf, 1996), including disparity models (Fleet et al., 1991; Rodrigues and du Buf, 2004a), have become available. On the basis of such models and neural processing schemes, it is possible to create a cortical architecture for figure-ground segregation (Hupe et al., 2001; Rodrigues and du Buf, 2006) and visual attention or Focus-of-Attention (FoA) (Parkhurst et al., 2002; Rodrigues and du Buf, 2005a). In addition, object detection, categorisation and recognition can be obtained by means of bottom-up and top-down data streams in the so-called "what" and "where" subsystems (Rensink, 2000; Deco and Rolls, 2004; Rodrigues and du Buf, 2006).

We will focus exclusively on keypoints in this paper. Heitger et al. (1992) developed a single-scale basis model that consists of single and double end-stopped cells in combination with complex inhibition schemes. Würtz and Lourens (1997) and Rodrigues and du Buf (2004) presented a pseudo-multi-scale approach, in which detection stabilisation at a fine scale is obtained by averaging keypoint positions over a few neighbouring, coarser, micro-scales. A truly multi-scale analysis was introduced later (Rodrigues and du Buf, 2005a). This idea was based on the fact that there are simple and complex cells tuned to different spatial frequencies, spanning multiple octaves; therefore, it can be expected that also end-stopped cells exist at all frequencies. We analysed the multi-scale keypoint representation, from very fine to very coarse scales, in order to study its importance and possibilities for developing a cortical architecture, with an emphasis on FoA. Also, a new aspect was included, namely the application of non-classical receptive-field (NCRF) inhibition to keypoint detection. Before, NCRF inhibition had only been applied to contour detection (Grigorescu et al., 2003), in order to separate object structures from surface textures. Below, we will argue that NCRF inhibition can be applied to edges *and* to keypoints, for creating two data streams dedicated to object structures and surface textures, but *only* at the finest scales. Furthermore, we will show that the multi-scale keypoint representation can be combined with automatic scale selection, for obtaining keypoints which are most characteristic of objects, and that it can play a role in object segregation. The latter two processes are thought to be essential in the what and where subsystems, for a rapid detection of where an object may be and a first categorisation to select most likely object templates in memory, after which all available features are used in object recognition.

A difficult and still challenging application, even in computer vision, is face detection. Despite the impressive number of methods devised for faces and facial landmarks (Yang et al., 2002), complicating factors are pose (frontal vs. profile), beards, moustaches and glasses, facial expression and image conditions (lighting, resolution). Despite these complications, we will study the multi-scale keypoint representation in the context of a plausible architecture for face detection. We add that we will not employ the multi-scale line/edge representation that also exists in area V1, in order to emphasise the importance of the information provided by keypoints. Also, we will not solve all complications referred to above, because we will argue, in the final Discussion, that low-level processing in area V1 needs to be embedded into a much wider context, including object templates stored in short- and long-term memory, and this context is expected to solve many problems.

There exists a vast literature concerning keypoints in computer vision, from basic feature extraction to object recognition, but much less in biological vision. Here we summarise a few approaches. Lourens and Würtz (1997) presented an object recognition system based on symbolic graphs, in which object corners are nodes and object contours are edges of the graphs. Their algorithm for corner detection is based on Heitger et al.'s (1992) model of cortical end-stopped cells, but they combined several scales and generalised to colour channels (Würtz and Lourens, 2000). Resulting corner detection was shown to be very stable in the presence of high-frequency textures, noise, varying contrast and rounded corners, see also Lourens et al. (2001) and Würtz and Lourens (2003). In this processing, graph edges are constructed by following contours between corners, using local evidence from the multi-scale Gabor wavelet transform. Model matching is achieved by finding subgraph isomorphisms in global image graphs.

Rosenthaler et al. (1992) also presented an integrated framework for extracting edges and keypoints. This detection scheme is based on analysis of oriented energy channels by using differential geometry. Barth et al. (1998) proposed end-stopped operators based on iterative, non-linear centre-surround inhibition. Henricsson and Heitger (1994) showed that an independent representation of corner and junction features provides suitable stop conditions for an aggregation process which allows to divide contours into meaningful substrings. They demonstrated that the active role of corners and junctions in the linking of contours greatly reduces problems associated with purely edge-based methods.

Lindeberg (1998) presented a detailed study of the Gaussian-derivative scale-space representation that can be used for a variety of early visual tasks. Operations like feature detection, which includes keypoints, feature classification and shape computation can be directly expressed in terms of (non-linear) combinations of Gaussian derivatives at multiple scales. Hansen et al.

(2001) developed a functional model of intra-cortical, recurrent, long-range interactions in V1 and proposed that long-range connections implement a multi-purpose preprocessing mechanism for main vision tasks, namely contour enhancement and corner detection. Later, Hansen and Neumann (2002) compared detected junctions based on the recurrent long-range interactions to junctions as obtained by a purely feed-forward model of complex cells. They also compared with two widely-used junction-detection schemes in computer vision, which are based on Gaussian curvature and the structure tensor. Ruzon and Tomasi (2001) used colour distributions to detect edges, junctions and corners, whereas Kovesi (2003) described corner and edge detection on the basis of the phase-congruency model. Triggs (2004) demonstrated that keypoints detected by the Förstner-Harris method are very stable when changing the illumination.

In addition to all different views and ideas referred to above, we mention two special projects. The first has no biological background, whereas the second has some minor biological background. The SUSAN project (Smith and Brady, 1997) concerns an approach to edge and corner detection with structure-preserving noise reduction. Non-linear filtering is used to define which parts of the image are closely related to each individual pixel, where each pixel is associated to a local image region which has about the same intensity (pixel values). Feature detectors are based on the minimisation of these local image regions, and the noise-reduction method uses the regions as smoothing neighbourhoods. The SIFT project (Lowe, 2004) has seen many developments along the years, for instance the extraction of distinctive image features from scale-invariant keypoints. Distinctive, invariant image features can be used for a reliable matching of different views of an object or a scene.

Most methods presented above have no direct biological background, and those *with* a clear biological background (Heitger et al., 1992; Barth et al., 1998; Hansen et al., 2001) are limited to one, fine scale. The only exceptions are the papers by Lourens and Würtz referenced above, in which a few (fine) scales are used for keypoint stabilisation. Furthermore, many methods are concerned with low-level feature extraction, for example for solving problems related to edge detection by employing keypoints. Extracted features are then used for high-level object detection in images, for example. In this paper, we study keypoint scale space, from the finest to very coarse scales, and show that this space can be exploited in building biological—and computer—vision systems.

2 Basic cell models and NCRF inhibition

Gabor quadrature filters provide a model of cortical simple cells (Lee, 1996). In the spatial domain (x, y) they consist of a real cosine and an imaginary

sine, both with a Gaussian envelope. A receptive field (RF) is denoted by (see for example Grigorescu et al. (2003))

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

with $\tilde{x} = x\cos\theta + y\sin\theta$ and $\tilde{y} = y\cos\theta - x\sin\theta$, the aspect ratio $\gamma = 0.5$ and σ determines the size of the RF. The spatial frequency is $1/\lambda$, λ being the wavelength. For the bandwidth σ/λ we use 0.56, which yields a half-response width of one octave. The angle θ determines the orientation (we use 8 orientations), and φ the symmetry (0 or $\pi/2$). We can apply a linear scaling between $f_{\min}$ and $f_{\max}$ with hundreds of contiguous scales. Below, the scale 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.

Responses of even and odd simple cells, which correspond to real and imaginary parts of a Gabor filter, are obtained by convolving the input image with the RFs, and are denoted by $R^E_{s,i}(x,y)$ and $R^O_{s,i}(x,y)$, s being the scale, i the orientation ($\theta_i = i\pi/(N_\theta - 1)$) and N_θ the number of orientations (here 8). Responses of complex cells are then modelled by the modulus

$$C_{s,i}(x,y) = [\{R^E_{s,i}(x,y)\}^2 + \{R^O_{s,i}(x,y)\}^2]^{1/2}. \tag{2}$$

There are two types of end-stopped cells (Heitger et al., 1992), single (S) and double (D). If $[\cdot]^+$ denotes the suppression of negative values, and $\mathcal{C}_i = \cos\theta_i$ and $\mathcal{S}_i = \sin\theta_i$, then

$$S_{s,i}(x,y) = [C_{s,i}(x + d\mathcal{S}_{s,i}, y - d\mathcal{C}_{s,i}) - C_{s,i}(x - d\mathcal{S}_{s,i}, y + d\mathcal{C}_{s,i})]^+ \tag{3}$$

and

$$\begin{aligned} D_{s,i}(x,y) = \Big[& C_{s,i}(x,y) - \frac{1}{2}C_{s,i}(x + 2d\mathcal{S}_{s,i}, y - 2d\mathcal{C}_{s,i}) \\ & - \frac{1}{2}C_{s,i}(x - 2d\mathcal{S}_{s,i}, y + 2d\mathcal{C}_{s,i})\Big]^+ . \end{aligned} \tag{4}$$

The distance d is scaled linearly with the filter scale s (we use $d = 0.6s$). Figure 1 shows end-stopped responses at three scales in the case of the traffic-sign image shown in Fig. 2. These responses mark the triangle and arrow etc. at a fine scale, but at coarser scales they are very diffuse due to the size of the RFs. In the next step, all end-stopped responses along straight lines and

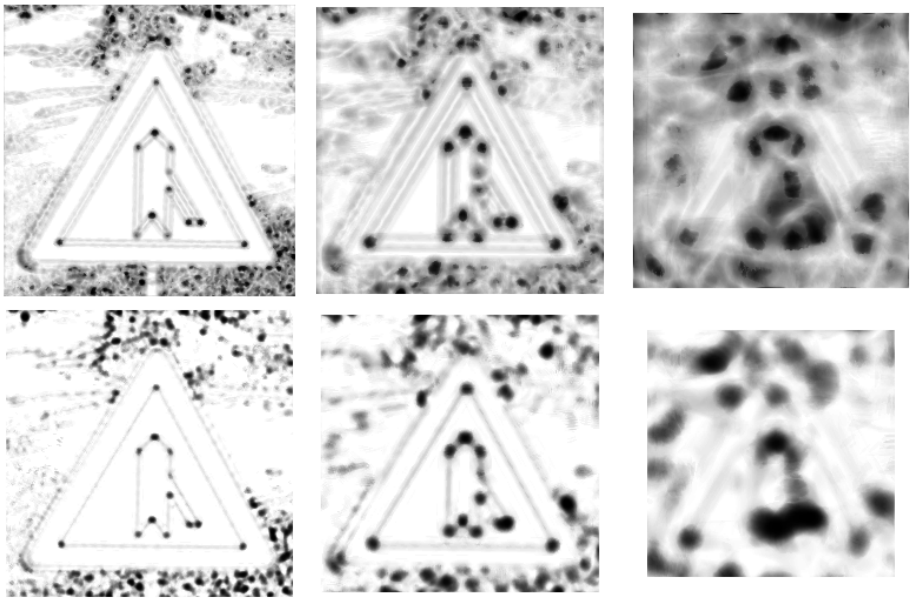

Fig. 1. Single (top) and double (bottom) end-stopped responses at three scales ($\lambda = 4, 8, 16$).

edges are suppressed, for which tangential (T) and radial (R) inhibition are used:

$$I_s^T(x,y) = \sum_{i=0}^{2N_\theta-1} [-C_{s,i \bmod N_\theta}(x,y) + C_{s,i \bmod N_\theta}(x + d\mathcal{C}_{s,i}, y + d\mathcal{S}_{s,i})]^+ \quad (5)$$

and

$$I_s^R(x,y) = \sum_{i=0}^{2N_\theta-1} \Big[C_{s,i \bmod N_\theta}(x,y) - 4 \cdot C_{s,(i+N_\theta/2) \bmod N_\theta}(x + \frac{d}{2}\mathcal{C}_{s,i}, y + \frac{d}{2}\mathcal{S}_{s,i})\Big]^+, \quad (6)$$

where $(i + N_\theta/2) \bmod N_\theta \perp i \bmod N_\theta$.

Non-classical receptive-field (NCRF) inhibition can be applied to suppress keypoints in textured regions. Models of NCRF inhibition are explained in more detail by Grigorescu et al. (2003). There are two inhibition types: (a) anisotropic, in which only responses obtained for the same preferred RF orientation contribute to the suppression, and (b) isotropic, in which all responses over all orientations contribute equally to the suppression.

The anisotropic NCRF (A-NCRF) model is computed by an inhibition term $t^A_{s,\sigma,i}$ for each orientation i, as a convolution of the complex cell responses $C_{s,i}$

with the weighting function w_σ, with

$$w_\sigma(x,y) = [\mathrm{DoG}_\sigma(x,y)]^+ / \|[\mathrm{DoG}_\sigma]^+\|_1, \tag{7}$$

where $\|\cdot\|_1$ is the L_1 norm and

$$\mathrm{DoG}_\sigma(x,y) = \frac{1}{2\pi(4\sigma)^2}\exp(-\frac{x^2+y^2}{2(4\sigma)^2}) - \frac{1}{2\pi\sigma^2}\exp(-\frac{x^2+y^2}{2\sigma^2}). \tag{8}$$

The operator $b^A_{s,\sigma,i}$ corresponds to the inhibition of $C_{s,i}$, i.e. $b^A_{s,\sigma,i} = [C_{s,i} - \alpha t^A_{s,\sigma,i}]^+$, with α controlling the strength of the inhibition.

The isotropic NCRF (I-NCRF) model is obtained by computing the inhibition term $t^I_{s,\sigma}$ which does not depend on orientation i. For this the maximum response map of the complex cells is constructed: $\tilde{C}_s = \max\{C_{s,i}\}$, with $i = 0, ...N_\theta - 1$. The isotropic inhibition term $t^I_{s,\sigma}$ is computed by the convolution of the maximum response map $\tilde{C}_s$ with the weighting function w_σ, and the isotropic operator is $b^I_{s,\sigma} = [\tilde{C}_s - \alpha t^I_{s,\sigma}]^+$.

3 Keypoint detection with NCRF inhibition at fine scale

As already mentioned, NCRF inhibition permits to suppress keypoints which are due to texture, for example in textured parts of an object surface. We experimented with the two types of NCRF inhibition introduced above, but here we only present the best results which were obtained by I-NCRF at the finest scale.

All responses of the end-stopped cells $S_s(x,y) = \sum_{i=0}^{N_\theta-1} S_{s,i}(x,y)$ and $D_s(x,y) = \sum_{i=0}^{N_\theta-1} D_{s,i}(x,y)$ are inhibited by $b^I_{s,\sigma}$, where $\alpha = 1$ is used, and we obtain the responses $\tilde{S}$ and $\tilde{D}$ of S and D that are above a small threshold of $b^I_{s,\sigma}$. Then we apply $I_s = I^T_s + I^R_s$ for obtaining the keypoint maps $\tilde{K}^S_s(x,y) = \tilde{S}_s(x,y) - gI_s(x,y)$ and $\tilde{K}^D_s(x,y) = \tilde{D}_s(x,y) - gI_s(x,y)$, with $g \approx 1.0$, and the final keypoint map $\tilde{K}_s(x,y) = \max\{\tilde{K}^S_s(x,y), \tilde{K}^D_s(x,y)\}$. In the last step, local maxima of $\tilde{K}_s(x,y)$ in x and y are detected.

Figure 2 shows, left to right, input images and keypoints detected at the finest scale that will be used in this paper, $\lambda = 4$, before and after I-NCRF inhibition. The face image ("face196") is part of the Psychological Image Collection at Stirling University (UK). As can be seen in Fig. 2, keypoint detection is very precise and mostly contour-related keypoints remain after inhibition. Although many texture-related keypoints have been suppressed, some may still appear because of strong, local contrast; see also Rodrigues and du Buf (2005a).

Fig. 2. Keypoints detected at the finest scale, without (centre) and with (right) NCRF inhibition.

Detected keypoints provide important image information because they code local image complexity, for example for FoA (see below), but we can go one step further. Object detection and recognition is helped much if detected positions are complemented by the type of complexity. In other words, it is useful to classify keypoints according to the underlying vertex structure, such as K, L, T, + etc. This is very difficult, because responses of simple and complex cells, which code the underlying lines and edges at the vertices, are unreliable due to response interference effects (du Buf, 1993). This implies that responses must be analysed in a larger neighbourhood around each keypoint. This problem has been solved by processing simple- and complex-cell responses in four cell layers, each layer comprising various grouping and detection cells. This process is very close to basic line and edge detection, see Rodrigues and du Buf (2004b), which is beyond the scope of this paper.

Figure 3 (left) shows two central, pentagonal, dendritic fields (shaded) and eight parallel ones around a keypoint, for directions 6 and 13. Grouping cells with such fields are necessary for probing simple and complex cells for dominant and sub-dominant *orientations* and then for symmetric or asymmetric *directions*. Figure 3 illustrates the application of keypoint classification to two traffic signs, at scale $\lambda = 4$. All keypoints of the "van" image have been detected, but three directions are still missing (encircled). There, structures have a size of 2 to 4 pixels, and we are at the very limit of what can be achieved by using Gabor filters. Also present in the "van" image is a keypoint (small diamond) that was detected near the top-right corner, but due to the lack of structure in its neighbourhood no direction has been attributed. In other

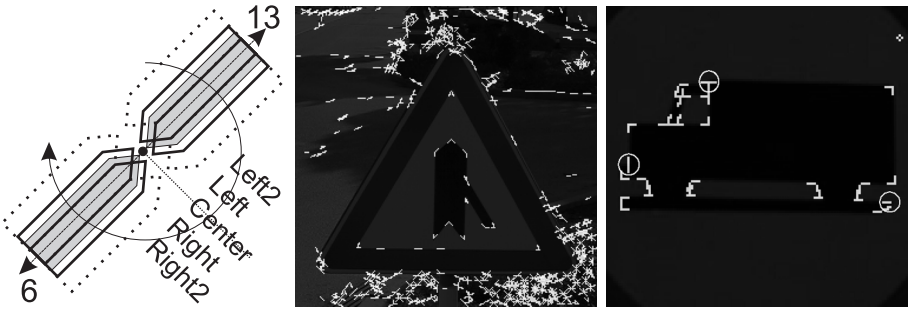

Fig. 3. Keypoint classification. Left: pentagonal dendritic fields of grouping cells that probe simple and complex cells for (sub)dominant orientations and (a)symmetric directions. Centre and right: detected keypoints with vertex structure.

words, this keypoint can be suppressed. It follows from Fig. 3 that detected and classified keypoints provide important information for object recognition, in this case the triangular sign with the arrow and the "van." This information must be complemented by lines and edges that are also extracted in area V1. Currently, the keypoint classification scheme is being implemented and optimised for application at arbitrary scale, but it is not yet clear whether vertex structure provides useful information in addition to detected lines and edges at coarse scales (Rodrigues and du Buf, 2004b).

4 Multi-scale keypoint representation

Although NCRF inhibition can be applied at any scale, we will not do this for two reasons: (a) we want to study keypoint behaviour in scale space for applications like FoA and facial landmark detection, and (b) in many cases a coarser scale, or increased RF size, will automatically eliminate keypoints in fine textures. In the multi-scale case keypoints are detected the same way as done above, but now by using $K_s^S(x,y) = S_s(x,y) - gI_s(x,y)$, $K_s^D(x,y) = D_s(x,y) - gI_s(x,y)$ and the final map $K_s(x,y) = \max\{K_s^S(x,y),\, K_s^D(x,y)\}$.

For analysing keypoint stability we can create an almost continuous, linear, scale space. In the case of Fig. 4, which shows projected trajectories of detected keypoints over scale in the case of a square and a star object, we applied 288 scales with $4 \leq \lambda \leq 40$. Figure 4 illustrates the general behaviour: at fine scales contour keypoints are detected, at coarser scales their trajectories converge, and at very coarse scales there is only one keypoint left near the centre of the object. However, it can also be seen (star object) that there are scale intervals where keypoints are unstable, even scales at which keypoints disappear and other scales at which they appear. (Dis)appearing keypoints are due to the size of the RFs in relation to the structure of the objects, analogous to Gaussian scale space (Koenderink, 1984; Lindeberg, 1994). Unstable

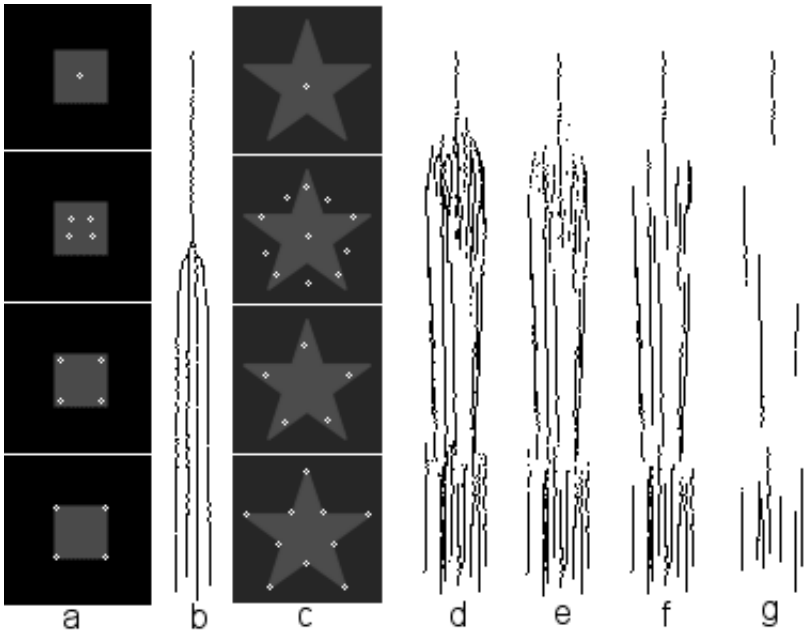

Fig. 4. Keypoint scale space, with finest scale at the bottom: (a) square, (b) projected 3D keypoint trajectories of square, (c) and (d) star and projected trajectories, (e) micro-scale stability, (f) and (g) stability over at least 10 and 40 scales, respectively.

keypoints can be eliminated by (a) requiring stability over a few neighbouring micro-scales (Rodrigues and du Buf, 2004b), by keeping keypoints that do not change position over 5 scales, the centre one and two above plus two below (Fig. 4e), or (b) requiring stability over at least N_s neighbouring scales (Figs 4f and 4g with $N_s = 10$ and 40, respectively). Such stabilisations are obtained by employing grouping cells with linear dendritic fields of different sizes over scale s. Assuming that keypoint cells are binary—they respond or they don't—grouping cells at all scales "sum" active keypoint cells, and if the sum (count) is below the necessary sum they can inhibit the keypoint cells. When keypoint cells may not be inhibited because of other processes, such as the ones described in the following sections, the grouping cells can inhibit gating cells which relay axons of keypoint cells.

The five leftmost columns in Fig. 5 illustrate that similar results are obtained after blurring, adding noise, rotation and rescaling of an object, a tree leaf, whereas the last two columns show results for other leaf shapes. In all cases, important contour keypoints remain at medium scales, and texture keypoints disappear without applying NCRF inhibition. In other words, NCRF inhibition is only useful for suppressing texture keypoints at the finest scales.

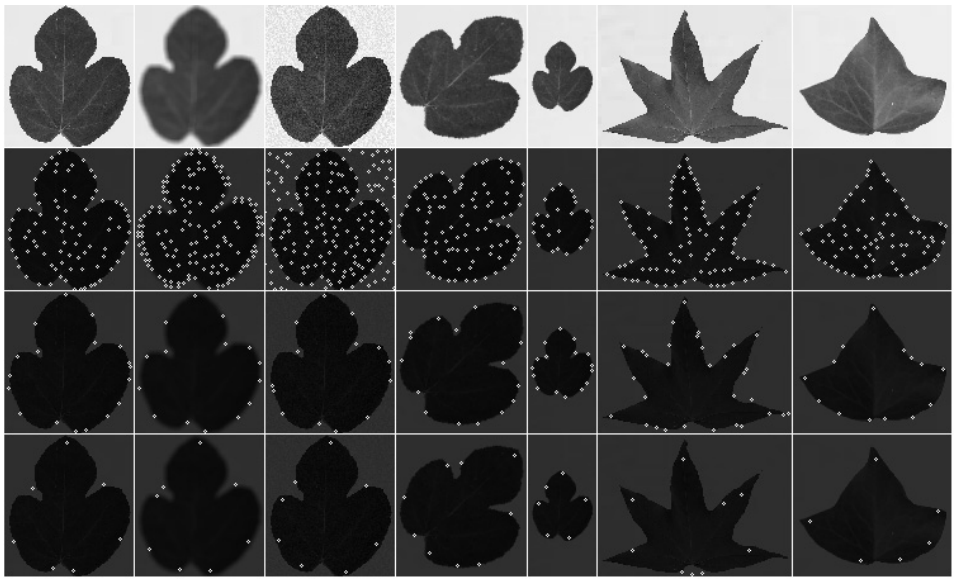

Fig. 5. From left to right: ideal image, blurred, with added noise, rotated and re-scaled leaf, plus two other leaves. Keypoints detected without NCRF inhibition, at fine (2nd line) and medium scales (bottom two lines).

5 Object segregation

The "bandwidth" of the what and where subsystems is very limited, because only one object can be attended at any time, which explains for example change blindness (Rensink, 2000). Both subsystems are "fed," bottom-up, by representations in area V1, and are "steered," top-down, from prefrontal (PF) cortex with templates of expected objects and expected positions (Deco and Rolls, 2004). The faster the bottom-up and top-down data streams converge, the faster an object will be detected and recognised. Typically, objects are recognised within 150–200 ms, and first category-specific activation of PF cortex starts after about 100 ms (Bar, 2003). This implies that some information propagates very rapidly from V1 to PF, such that the where system can select possible positions, after which the what system can test hypotheses. An important aspect in this is segregation, i.e., the separation of objects and the grouping of object features. Keypoints may play an important role in this process.

We have seen (Fig. 4) that keypoint trajectories converge from the contours at fine scales to the centres of objects at coarse scales. This implies that object segregation by means of a coarse-to-fine-scale strategy is feasible. Figure 6 (top) shows an image with four objects, two tree leaves, a star and the van from the traffic sign. Again, at very coarse scales the keypoints are located near the centres of the objects. In the case of the elongated van, an even coarser scale is required in order to obtain only one keypoint in the centre. Going from

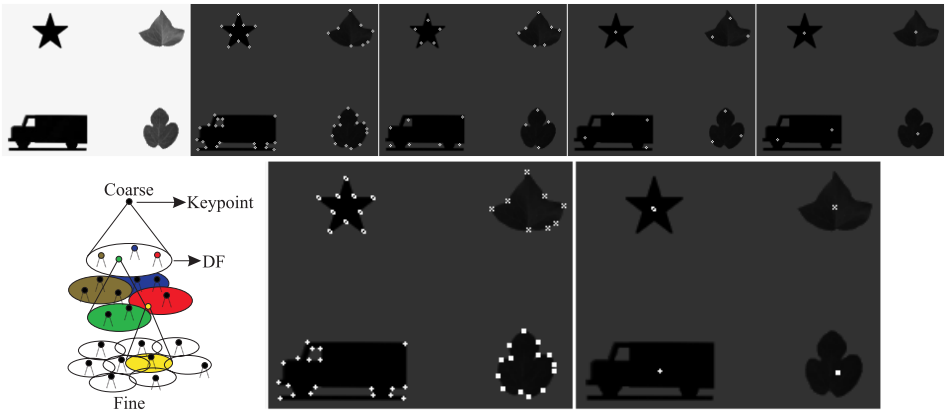

Fig. 6. Object segregation. Top: input image with four objects and detected keypoints at four scales ($4 \leq \lambda \leq 50$). Bottom: linking keypoints at a very coarse scale (right) to the finest scale (centre), with the principle (left; DF means dendritic field).

coarse to fine scales, keypoints will indicate more and more detail, until the finest scale is reached at which essential landmarks on contours remain.

At the coarsest level, each keypoint corresponds to one object. Each keypoint at a coarse scale is related to one or more keypoints at one finer scale, which can be slightly displaced. This relation is modelled by down-projection using grouping cells with a circular dendritic field, the size of which defines the region of influence. A responding keypoint cell activates a grouping cell. Only if the grouping cell is also excitated by responding keypoint cells one level lower, a grouping cell at the lower level is activated. This is repeated until the finest scale. Figure 6 illustrates the principle (bottom-left) with cones and the result (bottom-centre). The labels of the four keypoints at the coarsest scale, represented by different symbols, have been attributed to the keypoints at the finest scale. This coarse-to-fine-scale process permits to link all keypoints belonging to the same object. Results shown were obtained with $\lambda = [4, 50]$ and $\Delta\lambda = 4$.

A process as described above is supposed to occur completely in area V1, although information—including keypoints—at coarse scales propagates faster than information at fine scales to inferior-temporal (IT) cortex (Bar, 2003). This could imply that segregation is a dynamic effect or that it contributes dynamically to high-level object categorisation, which starts with coarse scales and is refined by adding finer scales. In any case, this process must be complemented by the what subsystem, because if two or more objects are very close, detected keypoints at very coarse scales will group the objects together and they can only be separated by probing specific object templates at finer scales. How this is done is not yet clear, because of feedback from higher areas like MT (Hupe et al., 2001), but it is done within 80 ms after image onset, which is late enough to allow contributions from higher visual areas (Zhaoping, 2003).

6 Automatic scale selection

Apart from object segregation, other processes may play an important role in the fast where and slower what subsystems. Concentrating on keypoints—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 keypoints which are most characteristic for an object are extracted and that these propagate first, for example for a rapid object categorisation. Above (Fig. 4) we have seen that different criteria for spatial stability over scales lead to different keypoint selections. One possibility is to select only one scale with the most characteristic keypoints. In computer vision, a similar approach has been applied by Lindeberg (1998), who selected the scale at which responses of Gaussian-derivative operators were strongest.

Here, we propose that the scale is the one at which the maximum number of *stable* keypoints is detected. This can be achieved with a few, simple processes, in which we assume again that outputs of keypoint cells are binary. First, a retinotopic map by means of grouping cells is created; see also below, i.e. saliency maps for FoA. A diagram of keypoint, grouping and gating cells is shown in Fig. 7. The grouping cells marked A have linear dendritic fields (solid black lines) that connect to keypoint cells (solid dots; active cells are big dots). These grouping cells sum all active keypoint cells at their position, over scale, which yields a sort of histogram. Second, at each scale, active keypoint cells activate gating cells (triangular synapses next to open circles). These cells 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. In other words, the latter grouping cells "count" stable keypoints at all individual scales. Fourth, the grouping cell with maximum activity is selected (winner takes all) and its axon activates other gating cells that gate outputs of keypoint cells at its scale. The outputs of the latter gating cells (Fig. 7, at top) provide the map which has the maximum number of stable keypoints. In the first step of this process, a scale-stability criterion as illustrated in Figs 4e, f or g can be included.

Figure 8 (top-left) shows the traffic-sign image with keypoints selected without applying a scale-stability criterion, which resembles detection at a fine scale after NCRF inhibition (Fig. 2, bottom-right). If stability over at least 20 scales is applied, many keypoints will disappear but the most important ones will remain (Fig. 8, top-centre and -right). In the case of the face image, important keypoints at eyes, nose, mouth and contour remain, even those at the marks on the forehead and cheekbone; compare Fig. 8 (top-right) with Fig. 2 (top-right). Also shown in Fig. 8 (bottom) are keypoints obtained with the SUSAN algorithm (Smith and Brady, 1997), the state-of-the-art, though limited to

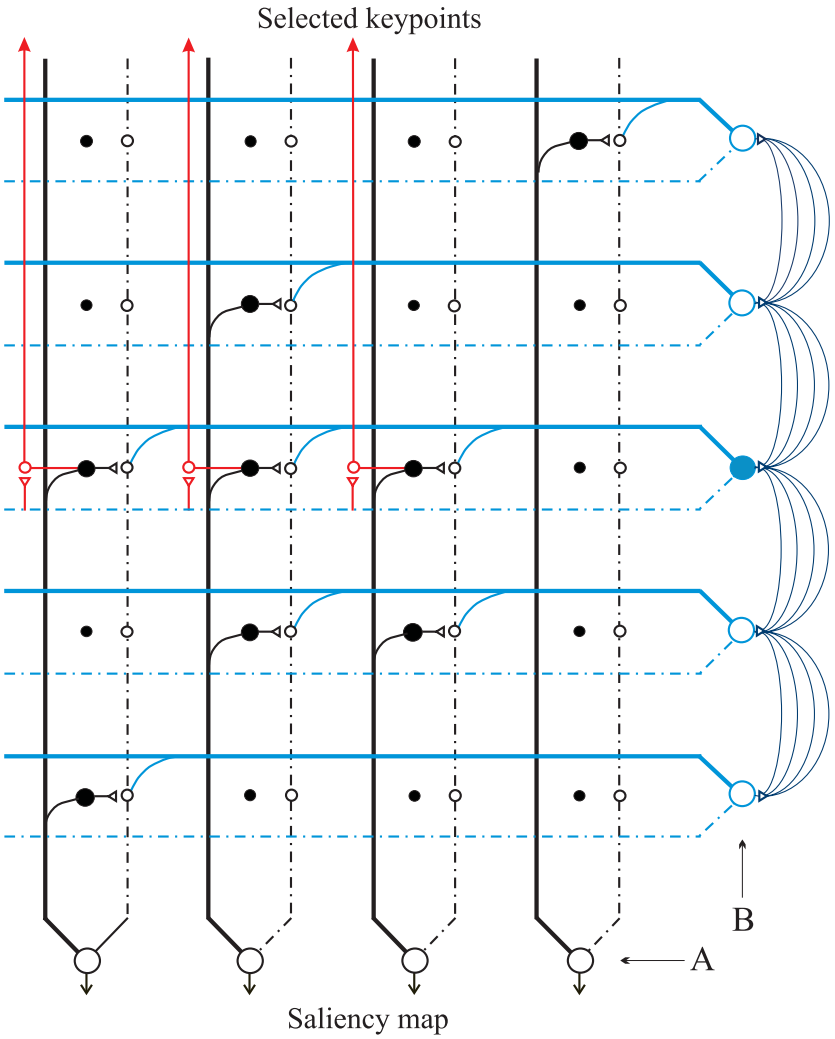

Fig. 7. Schematic diagram for automatic scale selection, with horizontally the position and vertically the scale. Keypoint cells are represented by solid dots (active keypoint cells by big dots), grouping cells by big, open circles, and gating cells by small open circles. Dendrites are shown by solid lines and axons by dash-dotted lines. See text.

fine scale, in computer vision. Comparing the SUSAN results with ours, we may conclude that advanced models of cortical processing can achieve similar, if not better, results.

7 Focus-of-Attention by saliency maps

As mentioned above, the what and where subsystems are steered, top-down, on the basis of expected objects and positions in PF cortex. However, there is one complication that has not yet been mentioned: our eyes are constantly mov-

Fig. 8. Top: results of automatic scale selection, without scale stability (left) and with stability over 20 scales (centre and right). Bottom: results obtained with the SUSAN algorithm (Smith and Brady, 1997).

ing in order to suppress static projections of blood vessels etc. in our retinas. During a fixation, stable information propagates from the retinas via the LGN to V1, where first features are extracted, and then, also during the next saccade, to higher areas. Fixation points in regions where complex—and therefore important—information can be found are much more important than points in homogeneous regions. Focus-of-Attention, for guiding the where system in parallel with steering our eyes, is thought to be driven by an attention component in PF cortex because of overt attention: while strongly fixating our eyes at one point, we can direct mental attention to points in the neighbourhood (Parkhurst et al., 2002). For modelling FoA we need a map, called saliency map, which indicates the most important points to be analysed (fixated). We propose a simple scheme based on the multi-scale keypoint representation, because keypoints code local image complexity.

As done in the previous section, activities of all keypoint cells at position (x, y) are summed over scale s by grouping cells. These cells are the ones marked A in Fig. 7. At positions where keypoints are stable over many scales, this summation map will show distinct peaks at centres of objects, important sub-structures and contour landmarks. The height of the peaks provides information about their relative importance. In addition, such a summation map, with some simple processing of the projected trajectories of unstable keypoints, like low-pass filtering and non-maximum suppression, might also contribute to solving the segregation problem: the object centre is linked to important structures, and these are linked to contour landmarks. Such a data

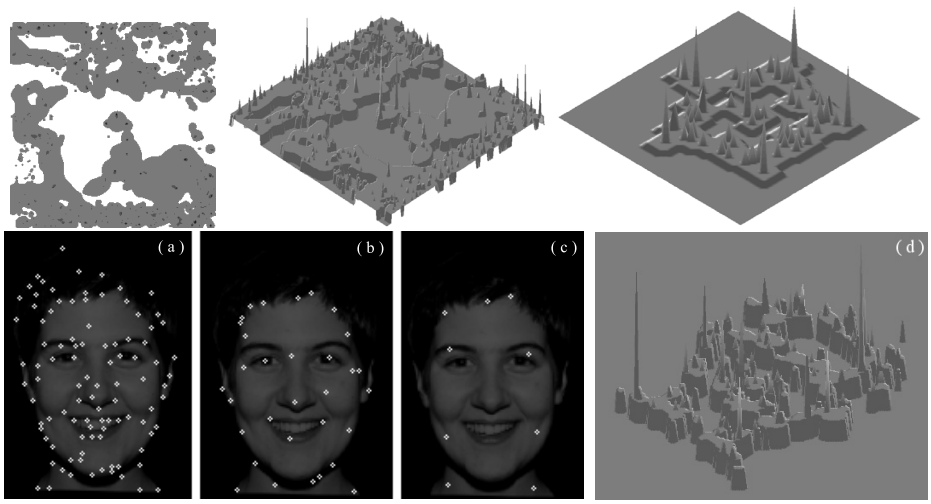

Fig. 9. Top, left to right: saliency maps in 2D and 3D of the traffic sign and star object. Bottom: keypoints at fine, medium and coarse scales, plus saliency map.

stream is data-driven and bottom-up, and could be combined with top-down processing from inferior-temporal cortex in order to actively probe the presence of objects in the visual field (Deco and Rolls, 2004). The summation map with links between the peaks might be available at higher cortical levels, where serial processing occurs for visual search, for example in the case when no object "pops out" and all objects must be screened sequentially.

Figure 9 (bottom; face196) shows keypoints at three different scales: (a) $\lambda = 4$, (b) $\lambda = 20$ and (c) $\lambda = 40$. We noticed that most if not all faces show a distinct keypoint on the middle of the line that connects the two eyes, like in Fig. 9b. Figure 9d shows the saliency map obtained on the basis of the entire scale space ($\lambda = [4, 40]$) with 288 scales. Important peaks are found at the eyes, nose and mouth, but also at the hairline and even the chin and neck. The regions around the peaks were created by a very simple process: each keypoint has a Region-of-Interest (RoI) that can be used to process—during a fixation—other information inside the RoI, such as lines, edges, textures and disparity. The RoI is small at fine scales and big at coarse scales. This is modelled by assuming circular axonal fields of keypoint cells, of size 3×3 at the finest scale ($\lambda = 4$) with linear scaling towards coarser scales. This means that the grouping cells marked A in Fig. 7 receive more input, and that the saliency map becomes a more diffuse "landscape" but still with high peaks. The maps shown in Fig. 9 have been thresholded, but this was only done for better displaying the structure of the maps, such that 3D projected views are not cluttered. The top row of Fig. 9 shows saliency maps in the case of the traffic-sign image (Fig. 2) and the star object (Fig. 4). In the former we can see the asymmetric region created by the keypoints at the bottom of and at the thin bar right to the arrow, in the latter the pentagonal structure of the star with peaks at the convex and concave vertices of the contour, in the triangles

and in the centre.

In Fig. 9d we can see the regions where important features are located, but it is quite difficult to see which peaks correspond to important facial landmarks. On the other hand, looking at Fig. 9b it is easy to see that some keypoints correspond to landmarks that we pretend to find in the next section, in this study limited to eyes, nose and mouth, but there are many more keypoints and at other scales (Fig. 9c) they are detected at other structures. Presumably, the visual system can use one "global" saliency map in combination with "partial" ones obtained by summing keypoints over smaller scale intervals, or even keypoints at individual scales, in order to optimise detection. This process can be steered by higher brain areas, which may contain prototype object maps with expected patterns, with approximate distances of eyes, nose and mouth. This can be part of the fast "where" data stream. Actual steering may consist of excitation and inhibition by pre-wired connections in keypoint scale space. This can be modelled by assuming grouping and gating cells which combine keypoint cells in approximate areas and at certain scales.

8 Application: face detection

In our simulations we explored one possible scenario, see also Rodrigues and du Buf (2005b). We assume the existence of very few layers of grouping cells, with dendritic fields in partial saliency maps that combine keypoints in specific scale intervals. The top layer with "face" cells groups axons of "eyes" (plural!), "nose" and "mouth" grouping cells. The "eyes" cells group axons of pairs of "eye" cells. Only the "eye," "nose" and "mouth" cells connect to the saliency maps, the "face" and "eyes" cells do not. This scenario consists of detecting possible positions of eyes, linking two eyes, then two eyes plus nose, and finally two eyes plus nose plus mouth. This is done dynamically by activating synaptic connections in the partial saliency maps, going from coarse to fine scales. We note that we did not yet include characteristic keypoints at other positions, like the one on the middle of the line that connects the two eyes (Fig. 9b).

We experimented with 30 faces—with different sizes and expressions—of the Stirling set (Fig. 13), and we used 7 partial saliency maps, each covering 40 scales distributed over $\Delta\lambda = 5$, but the scale intervals were overlapping 20 scales. The finest scale was at $\lambda = 4$. Examples of partial saliency maps are shown in Figs 10d and 11 (left). The search process starts at the coarsest scale interval, because there are much less candidate eye positions than there are at the finest scale interval, especially when a face is seen against a complex background. This is simulated by a feedback loop that activates connections to finer scale intervals, until at least one eye candidate is detected.

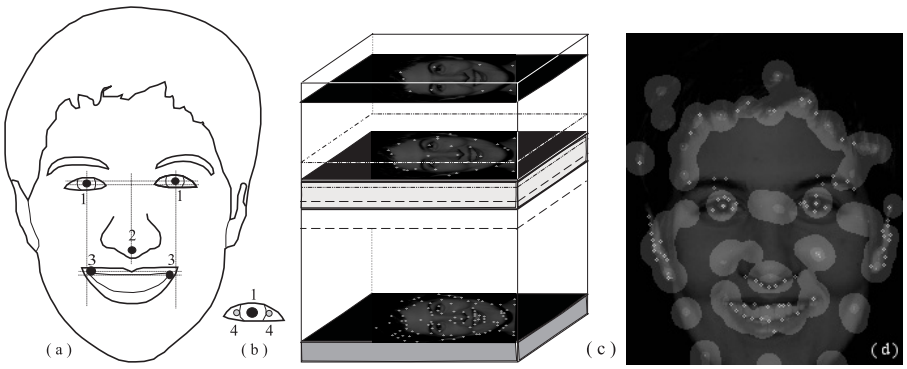

Fig. 10. Left to right: (a) facial landmarks, (b) eye landmarks, (c) impression of keypoint scale space, and (d) partial saliency map at fine scales ($\lambda = [4, 9]$) after NCRF inhibition.

First, "eye" cells respond to significant peaks (non-maximum suppression and thresholding) in the selected saliency map. In the case of "face196" this was the map at $\lambda = [13, 18]$, see Fig. 11 (left). This saliency map was the first one selected, because a peak at the centre of an eye, as indicated by Fig. 10b-1, may only be accepted if there also are two stable symmetric keypoints (eye corners) at the 40 finest scales ($\lambda = [4, 9]$), see Fig. 10b-4 and Fig. 10d. In order to reduce false positives, the latter is done after NCRF inhibition. If no single eye cell responds, the scale interval of the saliency map is not appropriate and the feedback loop will step through all saliency maps (Fig. 10c), until at least one eye cell responds.
Second, an "eyes" cell responds if two "eye" cells are active on an approximately horizontal line (Fig. 10a-1). An "eyes" cell is a grouping cell with two, symmetric, dendritic subfields. If no eye pair is detected, a new saliency map is selected (feedback loop).
Third, when two eyes can be grouped, a "nose" cell is activated, its dendritic field covering an area below the "eyes" and "eye" cells in the saliency map (Fig. 10a-2). If no peak is detected, a new saliency map is selected (feedback loop).
Fourth, if both "eyes" and "nose" cells respond, a "mouth" cell with two dendritic subfields at approximate positions of the two mouth corners (Fig. 10a-3) is activated. If keypoints are found, one "face" cell will be excited. If not, a new saliency map is selected (feedback loop).

The process stops when one or no face has been detected, but in reality it might continue at finer scale intervals because there may be more faces with different sizes in the visual field (image). The result obtained in the case of "face196" is shown in Fig. 11, where $+$, $\square$ and $\times$ symbols indicate detected and used keypoints at eyes, nose and mouth corners (actual positions of face and eyes cells are less important). More results are shown in Fig. 13. Of all 30 face images that we tested, one was problematic because of a very extreme

Fig. 11. Left: partial saliency map of face196 ($\lambda = [13, 18]$). Right: keypoints used by eye, nose and mouth detection cells.

expression, such that keypoints at mouth corners could not be grouped. In two cases, only the central part of a face was within the image border, which hampers detection of keypoints at eyes at coarse scales because of large filter sizes. In many applications such problems can be avoided. Nevertheless, a detection rate of 90% is encouraging in view of the extreme simplicity of the method, and compares well to other methods, which can be very complex and which must deal with the same problems (Yang et al., 2002).

Figure 13 also shows a correctly-detected face that was constructed by combining different fruits. The reason that this "fake" face was detected is that positions of facial landmarks as used in our model correspond to positions in real faces—our own visual impression, at first, also tells that it is a face. This effect is exploited in cartoons, even by the famous Italian painter Giuseppe Arcimboldo in the 16th century. Obviously, more features must be used, including the multi-scale line/edge representation in V1, but we must distinguish between face *detection* and *recognition*. Detection is thought to take place by means of keypoints and in the fast where system, after which additional features are available in the slower what system for recognition, including objects like fruits, in order to be able to distinguish between real and fake faces. In any case, we explored only one possible scenario in which grouping cells receive input at expected positions of eyes, nose and mouth. Such grouping cells might be located in V1, but also in V2 and V4, see Fig. 12 and Deco and Rolls' (2004) multi-area cortical architecture. As a consequence, only one "face cell" in V4 may be translation invariant, and therefore it may have a very large receptive field at the lowest (input) level. Figure 14 shows the result of applying our coarse-to-fine-scale scenario to an image with a complex background. This background leads to a huge number of keypoints, especially at the finest scales (Fig. 14 centre), with the possibility that random and unrelated keypoints can excitate "eye" and even "eyes" cells etc. However, this did not occur because of the coarse-to-fine strategy, in which a peak in the saliency map at a coarse scale (centre pupil) must be grouped with two keypoints at the finest scales

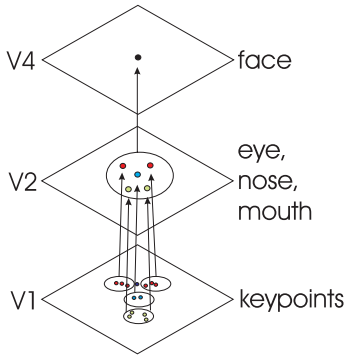

Fig. 12. Instead of grouping keypoints at facial landmarks in area V1, such groupings may actually be done in higher areas V2 and V4.

(eye corners). The additional groupings of keypoints at nose and mouth corners, at the coarser scales, increase selectivity. However, the result shown in Fig. 14 concerns a first experiment to test the detection scenario. Many more tests are required in order to validate and/or improve the method, including the detection of multiple faces—with different positions and sizes, eventually with partial occlusions—in images. This, and faces with different pose (frontal, 3/4 view, profile), requires the use of various templates in memory to steer the detection by activating different grouping cells with different spatial relations at different scales.

9 Discussion

As Rensink (2000) pointed out, the detailed and rich impression of our visual surround may not be caused by a rich representation in our visual memory, because the stable, physical surround already "acts" like memory. In addition, focussed attention is likely to deal with only one object at a time. His triadic architecture therefore separates focussed attention to coherent objects (System II) from non-attentional scene interpretation (Layout and Gist subsystems in System III), but both systems are fed by low-level feature detectors in System I.

In this paper we showed that keypoints detected on the basis of end-stopped operators, and in particular a few partial saliency maps that cover overlapping scale intervals, provide very important information for object detection. Exploring a very simple processing scheme, faces can be detected by grouping together axons of keypoint cells at approximate retinotopic positions, and this leads to robust detection in the case of different facial expressions. However, the simple scheme explored only works if the eyes are open, if the view is frontal, and if the faces are approximately vertical. For pose-, rotation- and

occlusion-invariant detection, the scheme must be fed by Rensink's short-term Layout and Gist subsystems, but also the long-term Scene Schema system that is supposed to build and store collections of object representations, for example of non-frontal faces.

We also showed that keypoints may play an important role in other cortical processes. A global saliency map provides ideal information for Focus-of-Attention, because distinct peaks are found at structures with a high complexity. This global saliency map can also be used for automatic scale selection, such that stable keypoints which are most characteristic for an object can be prepared for a first—but very fast—categorisation. Furthermore, it was shown that linking keypoints from coarse to fine scales can contribute to object segregation.

We focussed on the keypoint scale space in this paper. However, keypoint detection can be complemented with multi-scale line and edge detection, which is also supposed to occur in V1. It has already been shown that object segregation and categorisation—for example for distinguishing dogs, horses and cows—can also be achieved by only considering the line/edge scale space (Rodrigues and du Buf, 2006). This implies that the combination of detected keypoints and detected lines and edges will lead to improved performance, also enabling face *recognition*, but *how* all information can be combined in the best way remains an open question.

Owing to the impressive performance of current computers, it is now possible to test Rensink's (2000) triadic model in terms of Deco and Rolls' (2004) cortical architecture. The ventral what data stream (V1, V2, V4, IT) is supposed to be involved in object recognition, independently of position and scaling. The dorsal where stream (V1, V2, MT, PP) is responsible for maintaining a spatial map of an object's location, the spatial relationship of an object's parts, as well as moving the spatial allocation of attention. Both data streams are bottom-up and top-down. Apart from input via V1, both streams receive top-down input from a postulated short-term memory for shape features or templates in prefrontal cortical area 46, i.e., the more ventral area PF46v generates an object-based attentional component, whereas the more dorsal area PF46d specifies the location. As for now, we do not know *how* PF46 works. It might be the neurophysiological equivalent of the cognitive Scene Schema system mentioned above, but apparently the what and where data streams are necessary for obtaining view-independent object detection through cells with receptive fields of 50 degrees or more (Deco and Rolls, 2004). However, instead of receiving input directly from simple cells, the data streams should receive input from feature extraction engines in V1 and beyond, including keypoint cells!

Face images used are from the Psychological Image Collection at Stirling University (http://pics.psych.stir.ac.uk/). The SUSAN implementation is available at `http://www.fmrib.ox.ac.uk/~steve/susan/`. This research is partly financed by PRODEP III Medida 5, Action 5.3, and by the FCT program POSI, framework QCA III. We thank the anonymous reviewer for his comments that helped to improve the manuscript.

References

Bar, M., 2003. A cortical mechanism for triggering top-down facilitation in visual object recognition. J. Cogn. Neuroscience (15), 600–609.

Barth, E., Zetzsche, C., Krieger, G., 1998. Endstopped operators based on iterated nonlinear center-surround inhibition. Proc. Human Vision and Electronic Imaging III, SPIE, Vol. 3299, 67–78.

Bruce, V., Green, P., Georgeson, M., 2000. Visual Perception - Physiology, Psychology, and Ecology. Psychology Press Ltd (U.K.).

Deco, G., Rolls, E., 2004. A neurodynamical cortical model of visual attention and invariant object recognition. Vision Res. (44), 621–642.

du Buf, J., 1993. Responses of simple cells: events, interferences, and ambiguities. Biol. Cybern. 68, 321–333.

Elder, J., Sachs, A., 2004. Psychophysical receptive fields of edge detection mechanisms. Vision Res. 44, 795813.

Fleet, D., Jepson, A., Jenkin, M., 1991. Phase-based disparity measurement. CVGIP: Image Understanding 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.

Hansen, T., Neumann, H., 2002. A biologically motivated scheme for robust junction detection. Proc. 2nd Int. Worksh. Biologically Motivated Computer Vision, Springer LNCS 2525, 16–26.

Hansen, T., Sepp, W., Neumann, H., 2001. Recurrent long-range interactions in early vision. Springer LNCS 2036, 127–138.

Heitger, F., Rosenthaler, L., von der Heydt, R., Peterhans, E., Kübler, O., 1992. Simulation of neural contour mechanisms: from simple to end-stopped cells. Vision Res. 32 (5), 963–981.

Henricsson, O., Heitger, F., 1994. The role of key-points in finding contours. Proc. 3rd Europ. Conf. Computer Vision (Vol. II), Springer LNCS Vol. 801, 371–382.

Hubel, D., 1995. Eye, brain and vision. Scientific American Library.

Hupe, J., James, A., Girard, P., Lomber, S., Payne, B., Bullier, J., 2001.

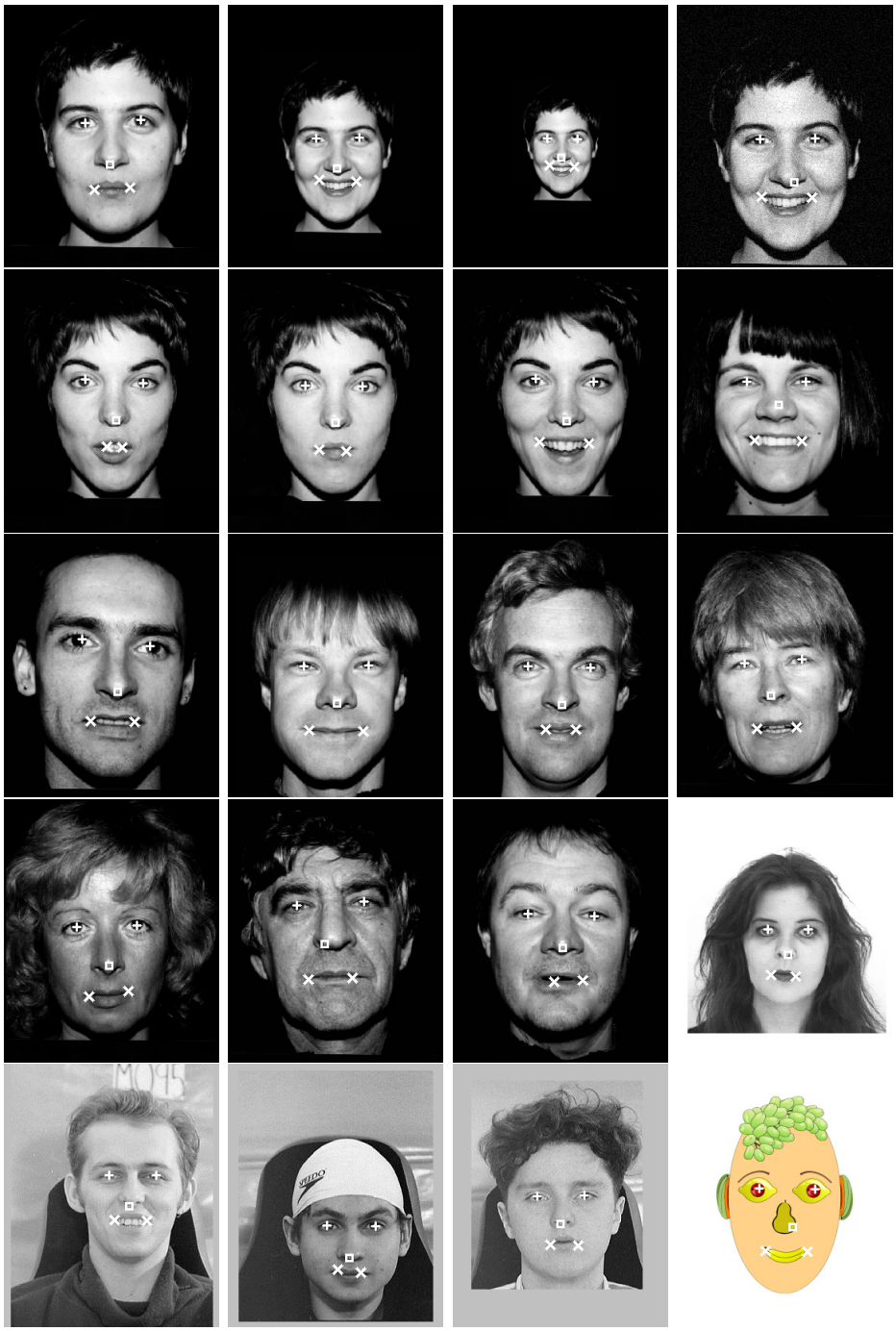

Fig. 13. Results obtained with different faces and expressions.

Feedback connections act on the early part of the responses in monkey visual cortex. J. Neurophysiol. 85 (1), 134–144.

Koenderink, J., 1984. The structure of images. Biol. Cybern. 50 (5), 363–370.

Kovesi, P., 2003. Phase congruency detects corners and edges. Proc. Australian Patt. Recogn. Soc. Conf., 309–318.

Lee, T., 1996. Image representation using 2D Gabor wavelets. IEEE Tr. PAMI

Fig. 14. Result with a complex background, which yields a huge number of keypoints especially at fine scales (centre).

18 (10), 959–971.

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

Lindeberg, T., 1998. Feature detection with automatic scale selection. Int. J. Computer Vision 30 (2), 77–116.

Lourens, T., Nakadai, K., Okuno, H., Kitano, H., 2001. Automatic graph extraction from color images. Proc. Europ. Symp. Artif. Neural Networks, 329–334.

Lourens, T., Würtz, R., 1997. Object recognition by matching symbolic edge graphs. Proc. 3rd Asian Conf. Computer Vision-Vol. II. Springer LNCS Vol. 1352, 193–200.

Lowe, D., 2004. Distinctive image feature from scale-invariant keypoints. Int. J. Comp. Vision 2 (60), 91–110.

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.

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

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

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

Rodrigues, J., du Buf, J., 2004a. Vision frontend with a new disparity model. Early Cogn. Vision Worksh., Isle of Skye (Scotland) www.cn.stir.ac.uk/ecovision-ws/.

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

Rodrigues, J., du Buf, J., 2005a. Multi-scale cortical keypoint representation for attention and object detection. Proc. 2nd Iberian Conf. on Patt. Recogn. and Image Anal. Springer LNCS Vol. 3523, 255–262.

Rodrigues, J., du Buf, J., 2005b. Multi-scale keypoints in V1 and face detec-

tion. Proc. 1st Int. Symp. Brain, Vision and Artif. Intell., Naples (Italy), Springer LNCS Vol. 3704, 205–214.

Rodrigues, J., du Buf, J., 2006. Cortical object segregation and categorization by multi-scale line and edge coding. Accepted for: Int. Conf. Computer Vision, Theory and Applications, 25-28 Febr. Setúbal, Portugal.

Rosenthaler, L., Heitger, F., Kübler, O., von der Heydt, R., 1992. Detection of general edges and keypoints. Proc. 2nd Europ. Conf. Computer Vision Springer LNCS Vol. 588, 78–86.

Ruzon, M., Tomasi, C., 2001. Edge, junction, and corner detection using color distributions. IEEE Tr. PAMI 23 (11), 1281–1295.

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

Triggs, B., 2004. Detecting keypoints with stable position, orientation and scale under illumination changes. Proc. Europ. Conf. Computer Vision IV, 100–113.

van Deemter, J., du Buf, J., 1996. Simultaneous detection of lines and edges using compound Gabor filters. Int. J. Patt. Recogn. Artif. Intell. 14 (6), 757–777.

Würtz, R., Lourens, T., 2000. Corner detection in color images by multiscale combination of end-stopped cortical cells. Image and Vision Comp. 18 (6-7), 531–541.

Würtz, R., Lourens, T., 2003. Extraction and matchimg of symbolic contour graphs. Int. J. Patt. Recogn. Artif. Intell. 17 (7), 1279–1302.

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

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