Springer

Dear Author,

Here are the proofs of your article.

- You can submit your corrections **online** or by **fax**.
- For **online** submission please insert your corrections in the online correction form. Always indicate the line number to which the correction refers.
- Please return your proof together with the **permission to publish** confirmation.
- For **fax** submission, please ensure that your corrections are clearly legible. Use a fine black pen and write the correction in the margin, not too close to the edge of the page.
- Remember to note the journal title, article number, and your name when sending your response via e-mail, fax or regular mail.
- **Check** the metadata sheet to make sure that the header information, especially author names and the corresponding affiliations are correctly shown.
- **Check** the questions that may have arisen during copy editing and insert your answers/ corrections.
- **Check** that the text is complete and that all figures, tables and their legends are included. Also check the accuracy of special characters, equations, and electronic supplementary material if applicable. If necessary refer to the *Edited manuscript*.
- The publication of inaccurate data such as dosages and units can have serious consequences. Please take particular care that all such details are correct.
- Please **do not** make changes that involve only matters of style. We have generally introduced forms that follow the journal's style.
  Substantial changes in content, e.g., new results, corrected values, title and authorship are not allowed without the approval of the responsible editor. In such a case, please contact the Editorial Office and return his/her consent together with the proof.
- If we do not receive your corrections **within 48 hours**, we will send you a reminder.

Please note

Your article will be published **Online First** approximately one week after receipt of your corrected proofs. This is the **official first publication** citable with the DOI. **Further changes are, therefore, not possible**.

After online publication, subscribers (personal/institutional) to this journal will have access to the complete article via the DOI using the URL: http://dx.doi.org/[DOI].

If you would like to know when your article has been published online, take advantage of our free alert service. For registration and further information go to: www.springerlink.com.

Due to the electronic nature of the procedure, the manuscript and the original figures will only be returned to you on special request. When you return your corrections, please inform us, if you would like to have these documents returned.

The **printed version** will follow in a forthcoming issue.

Fax to: +44 870 622 1325 (UK) or +44 870 762 8807 (UK)

To: **Springer Correction Team**
6&7, 5th Street, Radhakrishnan Salai, Chennai, Tamil Nadu, India – 600004

Re: Cognitive Processing DOI:10.1007/s10339-009-0262-2
A cortical framework for invariant object categorization and recognition

Authors: João Rodrigues · J.M. Hans du Buf

Permission to publish

I have checked the proofs of my article and

- ❑ I have no corrections. The article is ready to be published without changes.
- ❑ I have a few corrections. I am enclosing the following pages:
- ❑ I have made many corrections. Enclosed is the complete article.

Date / signature _____

Metadata of the article that will be visualized in OnlineFirst

Please note: Images will appear in color online but will be printed in black and white.

ArticleTitle	A cortical framework for invariant object categorization and recognition	
Article Sub-Title		
Article CopyRight - Year	Marta Olivetti Belardinelli and Springer-Verlag 2009 (This will be the copyright line in the final PDF)	
Journal Name	Cognitive Processing	
Corresponding Author	Family Name	**Rodrigues**
	Particle	
	Given Name	**João**
	Suffix	
	Division	Vision Laboratory, Institute for Systems and Robotics (ISR)
	Organization	University of the Algarve
	Address	Campus de Gambelas, FCT, 8000-810, Faro, Portugal
	Division	Escola Superior de Tecnologia
	Organization	University of the Algarve
	Address	Campus da Penha, 8005-139, Faro, Portugal
	Email	jrodrig@ualg.pt
Author	Family Name	**Hans du Buf**
	Particle	
	Given Name	**J. M.**
	Suffix	
	Division	Vision Laboratory, Institute for Systems and Robotics (ISR)
	Organization	University of the Algarve
	Address	Campus de Gambelas, FCT, 8000-810, Faro, Portugal
	Email	dubuf@ualg.pt
Schedule	Received	12 February 2007
	Revised	
	Accepted	6 May 2009

Abstract

In this paper we present a new model for invariant object categorization and recognition. It is based on explicit multi-scale features: lines, edges and keypoints are extracted from responses of simple, complex and end-stopped cells in cortical area V1, and keypoints are used to construct saliency maps for Focus-of-Attention. The model is a functional but dichotomous one, because keypoints are employed to model the “where” data stream, with dynamic routing of features from V1 to higher areas to obtain translation, rotation and size invariance, whereas lines and edges are employed in the “what” stream for object categorization and recognition. Furthermore, both the “where” and “what” pathways are dynamic in that information at coarse scales is employed first, after which information at progressively finer scales is added in order to refine the processes, i.e., both the dynamic feature routing and the categorization level. The construction of group and object templates, which are thought to be available in the prefrontal cortex with “what” and “where” components in PF46d and PF46v, is also illustrated. The model was tested in the framework of an integrated and biologically plausible architecture.

Keywords (separated by '-')	Categorization - Recognition - Dynamic routing - Cortical architecture
Footnote Information	

Journal: 10339
Article: 262

Author Query Form

Please ensure you fill out your response to the queries raised below and return this form along with your corrections

Dear Author

During the process of typesetting your article, the following queries have arisen. Please check your typeset proof carefully against the queries listed below and mark the necessary changes either directly on the proof/online grid or in the 'Author's response' area provided below

Query	Details required	Author's response
1.	Kindly provide page range for the reference Tarr (2005).	
2.	Kindly check and confirm the edit in table 1.	

Cogn Process
DOI 10.1007/s10339-009-0262-2

RESEARCH REPORT

A cortical framework for invariant object categorization and recognition

João Rodrigues · J. M. Hans du Buf

Received: 12 February 2007 / Accepted: 6 May 2009

Abstract In this paper we present a new model for invariant object categorization and recognition. It is based on explicit multi-scale features: lines, edges and keypoints are extracted from responses of simple, complex and end-stopped cells in cortical area V1, and keypoints are used to construct saliency maps for Focus-of-Attention. The model is a functional but dichotomous one, because keypoints are employed to model the "where" data stream, with dynamic routing of features from V1 to higher areas to obtain translation, rotation and size invariance, whereas lines and edges are employed in the "what" stream for object categorization and recognition. Furthermore, both the "where" and "what" pathways are dynamic in that information at coarse scales is employed first, after which information at progressively finer scales is added in order to refine the processes, i.e., both the dynamic feature routing and the categorization level. The construction of group and object templates, which are thought to be available in the prefrontal cortex with "what" and "where" components in PF46d and PF46v, is also illustrated. The model was tested in the framework of an integrated and biologically plausible architecture.

Keywords Categorization · Recognition · Dynamic routing · Cortical architecture

J. Rodrigues · J. M. Hans du Buf
Vision Laboratory, Institute for Systems and Robotics (ISR), University of the Algarve, Campus de Gambelas, FCT, 8000-810 Faro, Portugal
e-mail: dubuf@ualg.pt

J. Rodrigues (✉)
Escola Superior de Tecnologia, University of the Algarve, Campus da Penha, 8005-139 Faro, Portugal
e-mail: jrodrig@ualg.pt

Introduction

Object detection, segregation, categorization, and recognition are linked processes which cannot be completely sequential; they must be done in parallel, at least partially, and therefore they are overlapping; Rensink (2000). These processes are achieved in the ventral "what" and dorsal "where" pathways, Deco and Rolls (2004), with bottom-up feature extractions in areas V1, V2, V4, and IT[1] (what) in parallel with top-down attention from PP via MT to V2 and V1 (where). The latter is steered by possible object templates in memory, i.e., in prefrontal cortex with a "what" component in PF46v and a "where" component in PF46d. The Deco and Rolls model can explain invariance and attention besides the facts that cells at higher cortical areas have bigger receptive fields and that they are coding more complex patterns. However, their model is based on responses of simple cells in V1, whereas we are aiming at functional feature extractions in V1 and beyond. Although many image and object features are represented implicitly by simple cells, we apply explicit feature extractions: multi-scale line, edge and keypoint representations on the basis of cortical simple, complex and end-stopped cells; Rodrigues and du Buf (2006, 2008). The ultimate goal is to integrate feature extractions into a cortical architecture.

We are studying three related problems: when, where and how does categorization take place. The "when" problem allows for two hypotheses. The easy one is to assume that categorization occurs after recognition; Riesenhuber and Poggio (2000): if specific neurons respond in the case of recognizing dog-1, dog-2, and dog-3, a grouping cell can combine all responses: a dog. This view is too

[1] IT, inferior temporal cortex; PP, posterior parietal; MT, middle temporal; PF prefrontal.

Journal : Large 10339 | Dispatch : 18-5-2009 | Pages : 19
Article No. : 262 | LE | TYPESET
MS Code : COGPRO-D-07-00012 | CP | DISK

simplistic, because the system must collect evidence for a specific object or object group in order to select possible templates in memory. For example, when we glance a portrait made by Arcimbaldo, the famous, sixteenth-century Italian painter, our first reaction is "a face!", but then follows "fruits?" and finally "the cheek is an apple!"

When categorization occurs before recognition, Grill-Spector and Kanwisher (2005), the "where" problem is, at least partly, solved: it must take place at a very high level, with access to object templates in memory, and just before recognition. In fact, recognition can be seen as a last categorization step. Therefore, the "how" problem can be solved by taking into account feature extractions in V1 and beyond and the propagation of features to higher cortical areas. During the past years, we concentrated on the extraction of low-level primitives: lines, edges and keypoints, all multi-scale, see, e.g., Rodrigues and du Buf (2004, 2006, 2008). We showed that keypoint scale space provides ideal information for constructing saliency maps for Focus-of-Attention (FoA), and that the grouping of keypoints at different scales is robust for face detection; Rodrigues and du Buf (2006). Therefore, keypoints and FoA are thought to provide major cornerstones for the "where" system. In parallel, we showed that the multi-scale line/edge representation provides ideal information for object and face recognition, Rodrigues and du Buf (2008), i.e., in the "what" system. However, detection in the fast "where" pathway (a face!) must be linked with categorization and recognition in the slower "what" pathway (whose face?). The balance between the use of lines/edges and keypoints in the two pathways is still an open question.

A less open question concerns the use of features detected at different scales: information at coarse scales propagates first to higher areas, after which information at progressively finer scales arrives there, Bar (2004). This probably implies that coarse-scale information is used for a first, fast, but rough categorization, after which categorization is refined using information at progressively finer scales until an object is recognized. Bar (2003) proposed that a first categorization is based on a lowpass-filtered image of the object, but a smeared blob lacks structure. In our own experiments, Rodrigues and du Buf (2008), we therefore applied a different approach: after segregation, the coarse-scale line/edge representation of the outline is used for pre-categorization, after which all information is used for final categorization and recognition.

Any 3D object can lead to an infinite number of different projected images on the retinae due to variations in position, distance, lighting, and other factors including rotation and deformation. The ability to identify objects despite all possible transformations is central to visual object recognition. However, this still is a poorly understood mechanism, Cox et al. (2005), and transform-tolerant recognition remains a major problem in the development of artificial vision systems. In our brain, transform-invariant object recognition is automatic and robust, but it ultimately depends on experience; Tarr (2005). Recent findings, e.g., Cox et al. (2005), even support the idea that visual representations in the brain are plastic and largely a product of our visual environment and that invariant object representations are not rigid nor finalized—they are continually evolving entities, ready to adapt to changes in the environment. This idea complicates the classical idea of static representations in which only two but related problems need to be solved: (1) partial invariance to reasonable transformations like 2D rotation in the case of any canonical object view, which is addressed in this paper and (2) the total number of (3D) canonical object views that must be stored in memory. However, also plasticity can be explored at the two levels, in this paper in the form of dynamic routing for obtaining partial invariance to reasonable transformations.

There are several approaches to biological object recognition. Here, we focus briefly on approaches which, to some degree, are related to our own approach and architecture. Olshausen et al. (1993) described a model that relies on a set of control neurons, which dynamically modify the synaptic strengths of intracortical connections such that information from a windowed region of the primary cortex is selectively routed to higher cortical areas. Local spatial relationships (i.e. topography) within the attentional window are preserved as information is routed through the cortex. This enables attended objects to be represented in higher areas within an object-centered reference frame that is position and size invariant. Olshausen et al. hypothesize that the pulvinar (at the posterior part of the thalamus) may provide the control signals for routing information through the cortex. In preattentive mode, the control neurons receive their input from a low-level "saliency map" representing potentially interesting regions of a scene. During the pattern-recognition phase, control neurons are driven by the interaction between top-down (memory) and bottom-up (retinal input) sources.

In Rensink's (2000) triadic architecture, early preattentive processes feed both an attentional system concerned with coherent objects, and a non-attentional system concerned with scene gist and spatial layout. Instead of operating sequentially, the latter two subsystems operate concurrently for providing a context that can guide the allocation of attention. In this view, attention is no longer a central gateway through which all information must pass, but just one system that operates concurrently with several other (sub)systems. Furthermore, a scene is experienced via

Journal : Large 10339 Dispatch : 18-5-2009 Pages : 19
Article No. : 262 LE TYPESET
MS Code : COGPRO-D-07-00012 CP DISK

a "virtual representation" in which object representations are formed in a "just-in-time" fashion, only existing as long as they are needed.

The laterally interconnected synergetically self-organizing map (LISSOM), Miikkulainen et al. (2005), consists of a "family" of computational models which aim to replicate the detailed development of the visual cortex. The model can explain invariant (only size and viewpoint) detection of objects like faces. Hamker (2005) presented a feature-based computational model for invariant (but only translation) object detection in complex backgrounds (natural scenes) driven by attention in V4 and IT.

The collaboration called "Detection and Recognition of Objects in the Visual Cortex" integrates effort at several laboratories, aiming at a quantitative, hierarchical recognition model. The integrated architecture, like our own, reflects the general organization of the visual cortex in a stack of layers from V1 to IT to PF cortex; Serre and Riesenhuber (2004). Walther et al. (2005) are extending the basic recognition model by integrating a saliency-based and essentially bottom-up attentional model.

Deco and Rolls (2004) presented an invariant model that incorporates feedback-biasing effects of top-down attentional mechanisms in a hierarchically organized set of cortical areas with convergent feed forward connectivity, reciprocal feedback connections and local area competition. The model displays space-based and object-based covert visual search by using attentional top-down feedback from either the PP or the IT cortical modules, with interactions between the ventral and dorsal data streams occurring in V1 and V2. Deco and Rolls (2005) described a computational framework and showed how an attentional state held in short-term memory in PF cortex can, by top-down processing, influence the ventral and dorsal data streams in different cortical areas. Stringer et al. (2006) showed that invariant object recognition can be based on spatio-temporal continuity (during object translation and rotation) with "continuous transformation (CT) learning," which operates by mapping spatially similar input patterns to the same postsynaptic neurons in a competitive neural network system.

The goal of this paper is to show that low-level processing in terms of multi-scale feature extractions (keypoints, lines and edges) can be extended to higher-level processing: invariance in object categorization and recognition. We present a new model for obtaining 2D translation, rotation and size invariance by dynamic mapping of saliency maps based on multi-scale keypoint information. In addition, we present an integrated architecture in which coarse-scale information is used for a first but rough categorization, after which additional information at finer scales is used to refine categorization until objects are identified. As a consequence, extended models can cover more cognitive aspects in the near future. For example, processes like the learning of new objects or new, unexpected views of known objects will become subject to explicit modeling.

The rest of this paper is organized as follows: the next section deals with multi-scale feature extraction: lines, edges and keypoints plus the construction of saliency maps. Invariant categorization and recognition by dynamic routing, the construction of group templates and experimental results are presented in Section "Invariant object categorization and recognition". Section "The creation of group templates" concerns an integrated cortical architecture for the invariant categorization and recognition model. In the "Discussion" Section we discuss our approach and lines for future research. Mathematical formulations of the methods are provided in Appendix.

Lines, edges, keypoints and saliency maps

In order to explain the object categorization/recognition model, it is necessary to illustrate how our visual system can reconstruct, more or less, the input image. Image reconstruction can be based on one lowpass filter plus a complete set of bandpass-wavelet filters, such that the frequency domain is evenly covered. This concept is the basis of many image coding schemes. It could also be used in the visual cortex because simple cells in V1 are often modeled by complex Gabor wavelets. These are bandpass filters, Heitger et al. (1992), and lowpass information can be available through special retinal ganglion cells with photoreceptive dendrites which are *not* (in)directly connected to rods and cones, the main photoreceptors; Berson (2003). Activities of all cells could be combined by summing them in one cell layer that would provide a reconstruction or brightness map. But this creates a paradox: it is necessary to create *yet another observer* of this map in our brain.

The solution is simple: instead of summing all cell activities, we can assume that the visual system extracts lines and edges from simple- and complex-cell responses, which is necessary for object recognition, and that responding "line cells" are interpreted symbolically by a Gaussian cross-profile which is coupled to the scale of the underlying simple and complex cells. "Edge cells" are interpreted similarly, but with a bipolar, Gaussian-truncated error function profile (Rodrigues and du Buf 2008).

Responses of even and odd simple cells, corresponding to the real and imaginary parts of a Gabor filter, are denoted by R_s^E and R_s^O, s being the scale given by λ, the wavelength of the Gabor filters, in pixels (we assume that all different cells in the model can exist at all pixel positions). Responses of complex cells are modeled by the modulus C_s. For a detailed formulae see Appendix.

Journal : Large 10339 Dispatch : 18-5-2009 Pages : 19
Article No. : 262 LE TYPESET
MS Code : COGPRO-D-07-00012 CP DISK

The basic scheme for line and edge detection is based on responses of simple cells: 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 four possibilities for positive and negative events. For an improved, detailed scheme see Rodrigues and du Buf (2008) and Section "Cell models and multi-scale feature extraction" in Appendix.

Figure 1 (top row) shows lines and edges detected at eight scales $\lambda = \{4; 8; 12; 16; 20; 24; 28; 32\}$. Different levels of gray, from white to black, are used to show the events: positive/negative lines and positive/negative edges, respectively. As can be seen in Fig. 1, at fine scales many small events have been detected, whereas at coarser scales more global structures remain that convey a "sketchy" impression. Similar representations can be obtained by other multi-scale approaches; Lindeberg (1994). The middle row shows, from left to right, the input image, lowpass information, symbolic line and edge interpretations at a fine and a coarse scale, and the reconstructed image (see Section "Reconstruction model" in Appendix). Summarizing, the multi-scale line/edge interpretation with unipolar line and bipolar edge cross-profiles allows reconstructing the input image, and exactly the same representation will be used in the object categorization/recognition process.

Another important part of the model is based on responses of end-stopped cells in V1, which are very fuzzy and require optimized inhibition processes in order to detect keypoints at singularities. Recently, the original, single-scale model by Heitger et al. (1992) has been further stabilized and extended to arbitrary scale, and the multi-scale keypoint representation has been used to detect facial landmarks and faces; Rodrigues and du Buf (2005). There are two types of end-stopped cells: single and double. Responses of these are denoted by S_s and D_s, which correspond to the first and second derivatives of the responses of complex cells C_s. A final keypoint map K_s at scale s is obtained by combining local maxima of responses of single and double end-stopped cells after applying tangential and radial inhibition; see Rodrigues and du Buf (2006) for details, also Section "Cell models and multi-scale feature extraction" in Appendix. The bottom row in Fig. 1 shows detected keypoints (white diamonds) at fine (left) and coarse (right) scales superimposed on the darkened input image (at the same scales as used in the top row).

A saliency map for "driving" FoA—for details see Rodrigues and du Buf (2006)—can be obtained by summing keypoints over all scales. This provides a retinotopic (neighborhood-preserving) projection by grouping cells, and regions surrounding the peaks can be created by assuming that each keypoint has a certain Region-of-Influence, the size of which is coupled to the scale (size) of the underlying simple and complex cells. Keypoints which are stable over many scales will result in large and distinct peaks: at centers of objects (coarse scales), at important sub-structures (medium scales) and at contour landmarks (fine scales). The height of the peaks provides information about their relative importance. In other words, since keypoints are related to local image complexity, such a saliency map (SM) provides information for directing attention to image regions which are worth to be

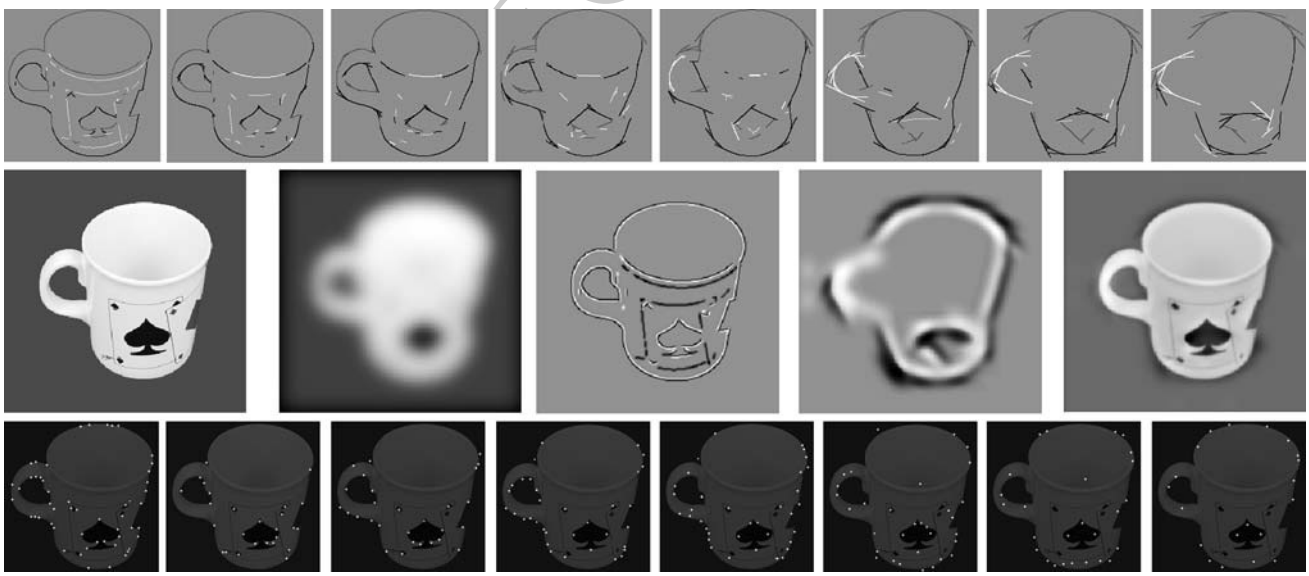

Fig. 1 *Top*: multi-scale line/edge detection in the case of a mug with, from left to right, fine to coarse scales. *Middle*: mug input image (*at left*) and reconstruction (*at right*) by combining lowpass information (second) and symbolic line/edge interpretations at a few scales (third and fourth images). Bottom: multi-scale keypoint representation of the mug with, from left to right, fine to coarse scales

Journal : Large 10339 | Dispatch : 18-5-2009 | Pages : 19
Article No. : 262 | ☐ LE | ☐ TYPESET
MS Code : COGPRO-D-07-00012 | ☑ CP | ☑ DISK

scrutinized, for example by steering the eyes in covert attention. This data stream is data-driven and bottom-up, and it can be combined with top-down processing from IT cortex in order to actively probe the presence of objects in the visual field; Deco and Rolls (2004). Examples of saliency maps can be seen in Figs. 4 and 5 (see also Section "Focus-of-Attention by saliency maps" in Appendix).

Before using the features in object recognition (next section) it makes sense to discuss whether they are robust enough against noise, i.e., whether they change position, appear or disappear after adding synthetic noise to object images or in real-world conditions under different illuminations and objects are seen against complex backgrounds. In our earlier experiments on multi-scale keypoints, Rodrigues and du Buf (2006), and on multi-scale lines and edges, Rodrigues and du Buf (2008), we showed that extracted features at coarse scales are very stable. However, at fine scales, especially at the finest scales with λ of the simple cells equal to a few pixels, significant changes are expected and they do occur. The same problem we encounter when trying to read a text which is too small: we automatically shorten the distance such that the text becomes bigger and our simple cells can resolve detail. In robot vision part of the solution is to measure the distance by the vergence of the stereo camera in combination with changing the zoom factor of the two lens systems. In our earlier experiments, Rodrigues and du Buf (2006, 2008), we explored other solutions for feature stabilization at fine scales: (a) non-classical receptive-field inhibition and (b) micro-scale stabilization. Micro-scale stabilization, i.e., keeping features that do not change in five of eight consecutive scales with $\Delta\lambda = 1$, proved to be the best method to apply to the entire scale space. In this case only a very few events may change position, but then only one pixel away. When measuring co-occurrences of features in input objects and templates stored in memory (next section) we therefore apply positional relaxation by using grouping cells with a certain dendritic field size, and micro-scale stabilization is applied to all features.

Invariant object categorization and recognition

To exemplify the model for invariant object categorization and recognition we selected eight groups of objects: dogs, horses, cows, apples, pears, tomatoes, cups, and cars, each with ten different images. The selected images were used at three levels: four types of objects (animals, fruits, cars, cups) for *pre-categorization*. Two of those were subdivided into three types (animals: horses, cows, dogs; fruits: tomatoes, pears, apples) for *categorization*. Final *recognition* concerns the identification of each individual object (e.g., horse number 3).

In our experiments we used the ETH-80 database, Leibe and Schiele (2003), in which all images are cropped such that they contain only one object, centered, against a 20% background. The views of all objects are also normalized, i.e., all animals with the head to the left (in Fig. 6 marked by white triangle). In order to test invariant processing, a set of modified input images was created by manipulations like translations, rotations, and zooms, including deformations (e.g., the head of a horse moved up or down relative to the body). We created 64 additional input images of the most distinct objects: 20 manipulated horse images (horses were used as a special test case for recognition); 6 dogs, 6 cows, 4 tomatoes, 4 pears and 4 apples, plus 10 cars and 10 cups. Figure 6 shows in neighboring left-right columns normalized objects and examples of modified objects. An exception is the top line which shows different manipulations: the normalized horse (marked by white triangle) with the head more down, bigger, and rotated and scaled against a white background. In what follows it is important to keep in mind that templates in memory are always based on original, normalized objects in the database, against which modified objects will be tested.

The creation of group templates

Good object templates in memory—both line/edge maps and saliency maps—are fundamental for obtaining good recognition results, but at the same time group templates must be generic enough to represent only one category for (pre-) categorization. Different line/edge templates with increasing detail are used in pre-categorization, categorization and final object recognition, but also different saliency maps in the dynamic routing for invariance (see next section).

The data structure of a template for each group has three components (at each scale): (a) the peaks of the saliency map (PSM), (b) the central keypoint (CKP) and (c) the line and edge information. For recognition and categorization we used the entire original and normalized objects, but for pre-categorization we used the segregated images extracted from the normalized objects (see Rodrigues and du Buf (2008) for how to extract the segregated information from the original image); Fig. 4c shows one example, a horse, in this case from an un-normalized image.

In order to create the group templates for pre-categorization (animal, fruit, car, cup), the saliency maps of the normalized objects in the database were selected randomly: for each group we summed half of the SMs, i.e., 5 SMs in the case of the 10 cups and cars, and 15 SMs in the case of animal (or fruit) with 10 images each of dogs, horses, and cows (or apples, pears and tomatoes). The resulting peaks

Journal : Large 10339 Dispatch : 18-5-2009 Pages : 19
Article No. : 262 LE TYPESET
MS Code : COGPRO-D-07-00012 CP DISK

(PSM) were obtained by non-maximum suppression and thresholding of the summed SMs. In case of the second categorization of animals and fruits, the same procedure was followed: five randomly selected SMs of horses, dogs and cows, and of apples, pears, and tomatoes. For final recognition the same procedure was used for each object individually, because we only have a single view of each object.

Essentially the same procedure was applied to the line/edge maps: random selections of images and logical combinations of event maps (for details see Rodrigues and du Buf (2008) and Section "Template data structure" in Appendix); binary events for pre-categorization (from the segregated images) and for categorization (from the original images), but considering events with type and polarity for final recognition.

It should be stressed that templates were always constructed on the basis of noise-free images. This is not to say that we expect serious problems, because micro-scale stabilization is applied to all features in combination with positional relaxation by grouping cells with a certain dendritic field size. As a matter of fact, group templates are always influenced by size and position variations due to approximate object normalizations: no two apples are exactly equal in size and position, nor are apples, pears, and tomatoes. Furthermore, our experimental results with local and global feature matching showed that sporadic feature variations are completely irrelevant. More relevant is the question how we can construct a system which is capable to construct (group) templates on the basis of unnormalized object views covering a certain size and viewpoint variation. By definition, such a system can cope with noise; both noise due to imaging conditions and to object variations. The goal in the near future is that the entire process will be implemented and tested in a completely dynamic way, including the integration of newly categorized or recognized objects into the (group) templates.

Rows 1 and 2 of Fig. 5 show the templates used in pre-categorization with, from left to right, saliency map, significant peaks and line/edge map at $\lambda = 32$ (one of three scales used) for the animal, fruit, car and cup groups. Rows 3 to 5 show the same for categorization ($\lambda = 8$ for the line/edge maps, one of eight scales used) with, from left to right: horse, cow, dog, tomato, pear and apple group templates. The bottom row shows two individual object templates used in recognition, i.e., two examples of the ten different horses, with the line/edge map at $\lambda = 4$ (one of eight scales used). In Summary, Fig. 5 shows the template information in memory on the basis of *normalized* objects against which *modified* objects will be matched. Appendix, Section "Template data structure" summarizes the template data structures.

Categorization and recognition by dynamic routing

For each object/template pair to be matched (categorization or recognition) a grouping cell, with its dendritic field (DF) in the SM, is positioned at the central keypoint (CKP) that represents the entire object/template at very coarse scales (Fig. 2a); this cell triggers the matching process (such central keypoints at coarse scales are always located at or close to an object's centroid; see Figs. 4 and 6 in Rodrigues and du Buf (2006)). The invariant method consists of steps a–f as follows:

(a) a) Central keypoints at very coarse scales of an input object and a template are made to coincide (Fig. 2b; T stands for translation). This can be seen as a translation of all keypoints (SM peaks) of the object to the ones of the template (or vice versa), but in reality there is no translation: only a dynamic routing by a hierarchy of grouping cells with DFs in intermediate neural layers such that the response of the central grouping cell of the template is maximum.

(b) The same routing principle of step (a) is applied to the two most significant SM peaks (from all scales), one of the input object and one of the template. Again, grouping cells at those peaks and with DFs in the

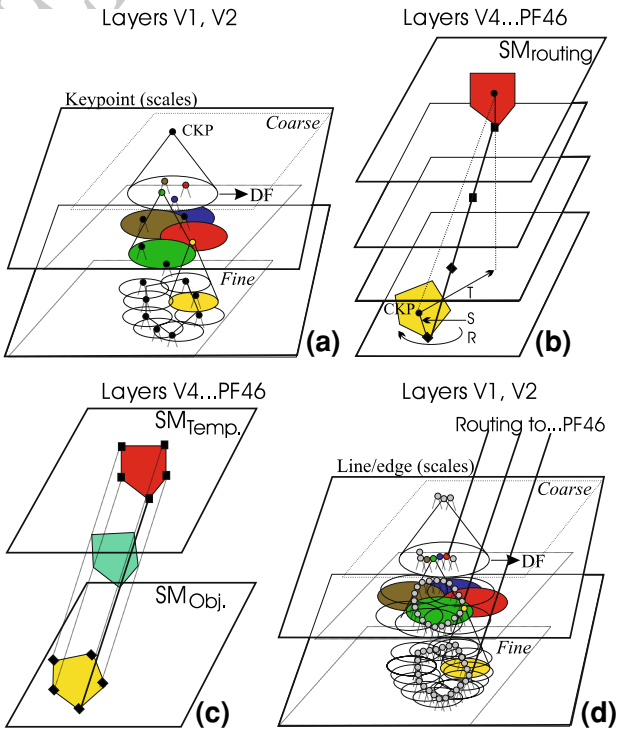

Fig. 2 Dynamic routing principle: **a** keypoints in scale space plus the central keypoint (CKP) at the coarsest scale, **b** the routing representation using the CKP and the highest saliency map peak (SMP) for the initial routing with translation (T), rotation (R) and scaling (S). The routing of saliency map peaks of an input object to those of a template in memory (**c**) is also applied to line edge events (**d**)

Journal : Large 10339 Dispatch : 18-5-2009 Pages : 19
Article No. : 262 LE TYPESET
MS Code : COGPRO-D-07-00012 CP DISK

intermediate layers serve to link the peaks by dynamic routing, but this time for compensating rotation and size (Fig. 2b; R and S). The resulting routing (translation, rotation and size projection) is then applied to all significant peaks (Fig. 2c) because they belong to a single object/template pair.

Figure 3 illustrates the above two steps. At top-left, central keypoints of template and input object excitate cells at intermediate levels through axonic fields, spreading activations in separate top-down (solid circle) and bottom-up (open circle) trees. This enables grouping cells at all levels to combine the top-down and bottom-up activations (shown in red). Once this first routing has been established, it is propagated laterally to routing cells at all levels. Using similar cell structures, most significant peaks in SMs are used to refine the routing (Fig. 3 top-right in green and bottom-left in blue). In the Discussion this process is also called "anchoring."

(c) All other significant SM peaks of the input object and the template are tested in order to check whether sufficient coinciding pairs exist for a match. To this end another hierarchy of grouping cells is used: from many local ones with a relatively small DF to cover small differences in position due to object deformations, etc., to one global one with a DF that covers the entire object/template. Instead of only summing activities in the DFs, these grouping cells can be inhibited if one input (peak amplitude of object, say)

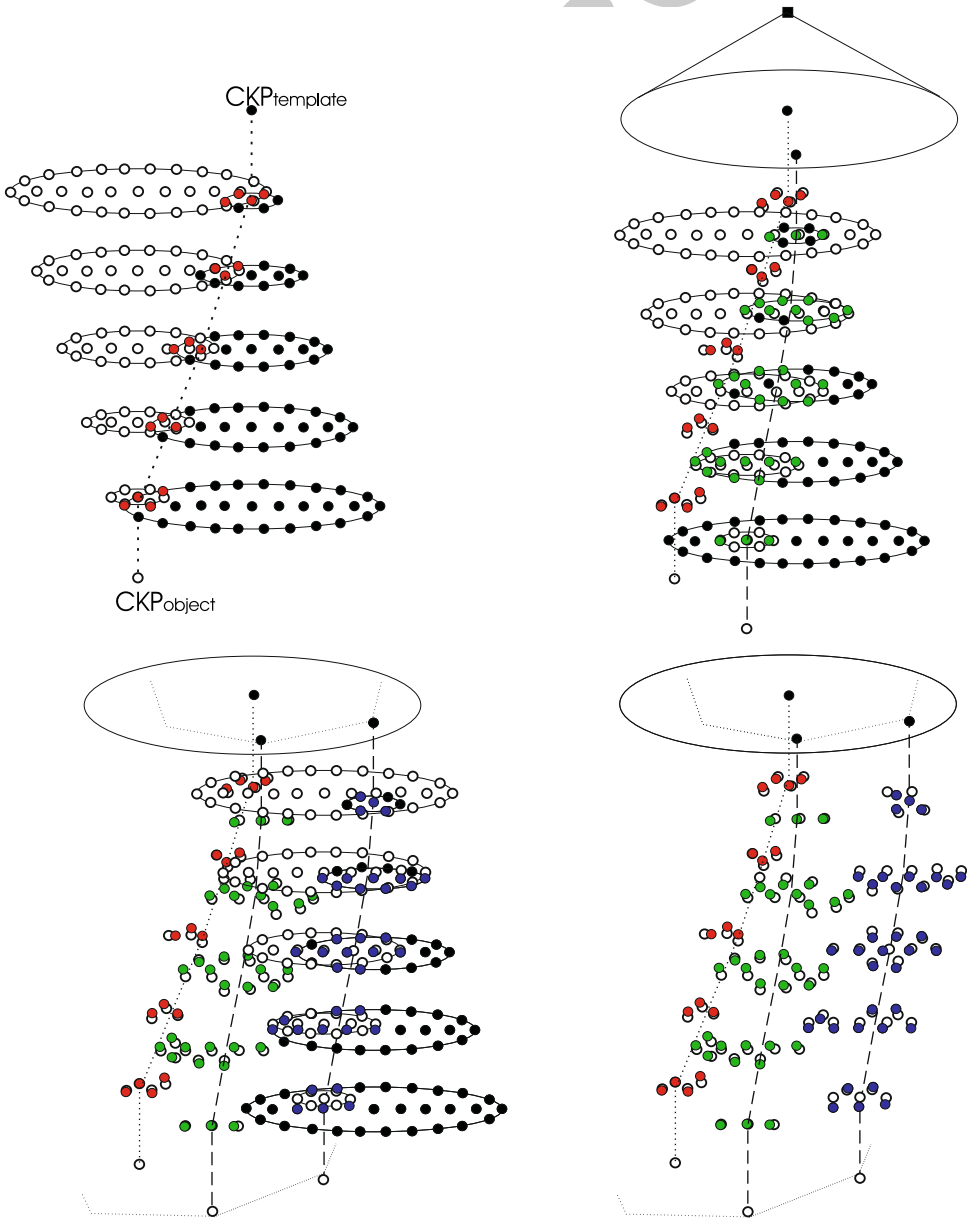

Fig. 3 Dynamic routing scheme with spreading and grouping: at *top-left*, central keypoints of template and input object excitate cells at intermediate levels through axonic fields, spreading activations in separate top-down (*solid circle*) and bottom-up (*open circle*) trees. *Top-right* and *bottom-left*: similar cell structures are used for the most significant peaks in SMs in order to refine the routing. *Bottom-right*: only the remaining activated cells are used for the routing

Journal : Large 10339 Dispatch : 18-5-2009 Pages : 19
Article No. : 262 LE TYPESET
MS Code : COGPRO-D-07-00012 CP DISK

is less than half of the other input (in this case of the template).

(d) If the global grouping of corresponding pairs of significant peaks is above a threshold (half of the maximum peak in the SM), the invariant match is positive. If not, this does not automatically mean that input object and template are different: the dynamic routing established in step (b) may be wrong. Steps (b-c) are then repeated by inhibiting the most significant peak of the object and selecting the next biggest peak.

(e) If no global match can be achieved, this means that the input object does not correspond to the template or that the view of the object (deformation, rotation or size) is not represented by the template. In this case the same processing is applied using all other templates in memory until the ones are found which could match. Although this process is simulated sequentially in our experiments, in reality this could be done in parallel by means of associative memory; Rehn and Sommer (2006).

(f) Up to here, only saliency maps were used to find possibly matching templates, but mainly for dynamic routing which virtually "superimposes" the input object and templates. In this step the dynamic routing of keypoints is also applied to the multi-scale line/edge representations in order to check whether an object and a template really correspond (Fig. 2d). Again, this is done by many grouping cells with small DFs (local correlation of line/edge events) and one with a big DF (global object/template correlation); see Rodrigues and du Buf (2008). The use of the small DFs can be seen as a relaxation: two edges of object and template count for a match if they are at the same position but also if they are very close to each other. The size of the DFs is coupled to the size of underlying complex cells.

The template information used in step (f) depends on the categorization level. In the case of the first, coarse, pre-categorization (f.1), only line/edge events (Fig. 4d) at three coarse scales of the segregated, binary object (Fig. 4c) are used, because (a) segregation must be done before categorization and (b) coarse-scale information propagates first from V1 to higher cortical areas; Bar et al. (2006). Global groupings of lines and edges are compared over all possibly matching templates, scale by scale, and then summed over the three scales, and the template with the maximum sum is selected (winner-takes-all). Figure 4f shows a projected and matching line/edge map after dynamic routing. In the case of the subsequent finer categorization (f.2), the process is similar, but now we use line/edge events at all eight scales obtained from the object itself instead of from the binary segregation. Figure 4g and h show projected peaks and the line/edge map used in categorization. Final recognition (f.3) differs from categorization (f.2) in that line and edge events are treated separately: object lines must match template lines and edges must match edges. This involves three additional layers of grouping cells, two for local co-occurrences of lines and edges and one global. Figure 4i and j show projected peaks and the line/edge map used in recognition. See Rodrigues and du Buf (2008) for complete explanations of the matching processes in the case of using only normalized object views, also

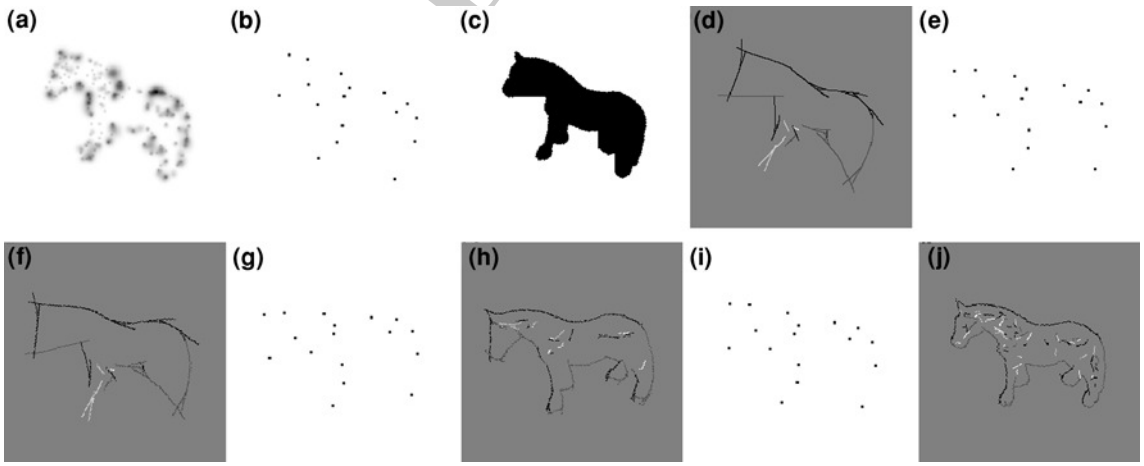

Fig. 4 Invariant categorization and recognition steps: **a** Saliency map of modified horse8, **b** SM peaks, **c** segregated object and **d** line/edge coding of segregated object at $\lambda = 24$. **e**, **f** SM peaks and line/edge map of normalized horse8 (after dynamic routing) in pre-categorization. **g**, **h** The same with line/edge map at $\lambda = 8$ in categorization. **i**, **j** The same with line/edge map at $\lambda = 4$ in final recognition. Input object and matching object (used only in recognition) are shown in Fig. 6 (marked by a *black* and *white corner triangle*)

Journal : Large 10339 Dispatch : 18-5-2009 Pages : 19
Article No. : 262 LE TYPESET
MS Code : COGPRO-D-07-00012 CP DISK

Appendix, Sections "Dynamic routing" and "Similarity between objects and templates".

Results

The results obtained were quite good: from the 64 modified input images, pre-categorization (animal, fruit, car, and cup) failed in 12 cases. Of the remaining 52 images, categorization (animal: horse, cow, dog; fruit: tomato, pear, apple) failed in 8 cases. Recognition failed for 4 of the 44 remaining images. The final recognition rate is therefore 62.5%. However, the above numbers are not definitive because they concern a first test of the concept and many errors can be explained. For example, some image manipulations were too extreme and we could have selected less extreme manipulations.

As for our previous results obtained with only normalized objects, Rodrigues and du Buf (2008), categorization errors occurred mainly for apples and tomatoes, which can be explained by the fact that the shapes are very similar and no color information has been used. In pre-categorization some fruits were categorized as cups. This mainly concerned pears and can be explained by the tapered-elliptical shape in combination with size variations, such that keypoints and line/edge events of input pears can coincide with those of the cups-group template (Fig. 5 top-right). As expected, especially in the case of recognition, problems occurred with extreme size variations. The scales used ($\lambda = [4, 32]$) are related to the size of the objects and the level of detail that can be represented. Figure 6 (middle three images in the fourth column) shows the smallest objects that could be dealt with by using these scales. The image at bottom-right proved too extreme (all modified objects shown on the bottom line were not correctly categorized or recognized).

It should be emphasized that the method can be applied to images which contain multiple objects. Although our visual system has a limited "bandwidth" and can test only one object at any time Rensink (2000), this problem can be solved by sequential processing of all detected and segregated objects. However, if object segregation and recognition are coupled processes, we are left with a typical chicken-or-egg problem, unless the process is controlled by, e.g., the gist system (see Discussion). Finally, it should be mentioned that dynamic routing of keypoints (significant peaks in saliency maps) and line/edge events in intermediate neural layers has consequences for the minimum number of canonical object views in memory, i.e., the number of templates. If a horse template has the head to the left and the legs down, but an input horse has been rotated (2D) by 180 degrees such that the head is to the right and the legs are up, dynamic routing will not be possible because there will be a crossing point in the routing at some neural layer. In this case a separate template is necessary. In addition, recognition in the case of 3D rotation may require more templates because of asymmetrical patterns of a horse's fell on its left and right flanks.

Integrating the architecture

The invariant object categorization and recognition model must be integrated into a cortical architecture, where the first task is to get the gist of the scene by a rapid but global classification; Oliva and Torralba (2006). After this all the objects can be analyzed, but sequentially, i.e., only one object at any time; Rensink (2000). Individual objects are analyzed in a multi-level recognition process, Grill-Spector and Kanwisher (2005), and interesting positions to be analyzed after the gist stage are stored in a "waiting list" (normally, this is modeled by sequential processing of most-to-less-important peaks in a saliency map, simulating eye movements and fixation points, with inhibition of returns to already analyzed positions; Prime and Ward (2006).

Objects can be categorized or recognized at different levels, and some objects do need several processing levels before recognition is achieved. For example, in the case of a horse called Ted recognition can be achieved after three levels: animal, horse, Ted. However, this is a very rigid scheme in which all horses need to go through all levels. If Ted's fell is very characteristic, and no other known object, animal or car. etc., displays a similar pattern, Ted could be recognized instantaneously by using other information channels, for example, devoted to color and/or texture. But such channels are not yet implemented and our model is restricted to multi-scale line/edge and keypoint representations. Nevertheless, also in our model an object can be recognized at an early level, if a measure for correspondence—a match with one template in memory—is much bigger than a threshold level and correspondence measures of all other templates are much smaller than the threshold.

Figure 7 shows in a "features and blocks" fashion the generalized architecture, where each block represents the type of feature involved (and scales), as well as the processing done at the different stages. The blocks are displayed in a sequential way with early processing at the top and later processing toward the bottom. Only three levels are shown (1, 2 and n), but n is variable. At each level, three templates are shown (A, B and N), but N is variable and a function of the level. Features are indicated by SM (saliency map), LE (line-edge code), and LE repr. (symbolic line/edge representation), the latter two with an indication of the scales used (All scales or LF meaning coarse scales only). The arrows show the information flow, the circles indicate activations, and dashed arrows represent feedback loops. If a template cannot reach a global

Journal : Large 10339 Dispatch : 18-5-2009 Pages : 19
Article No. : 262 LE TYPESET
MS Code : COGPRO-D-07-00012 CP DISK

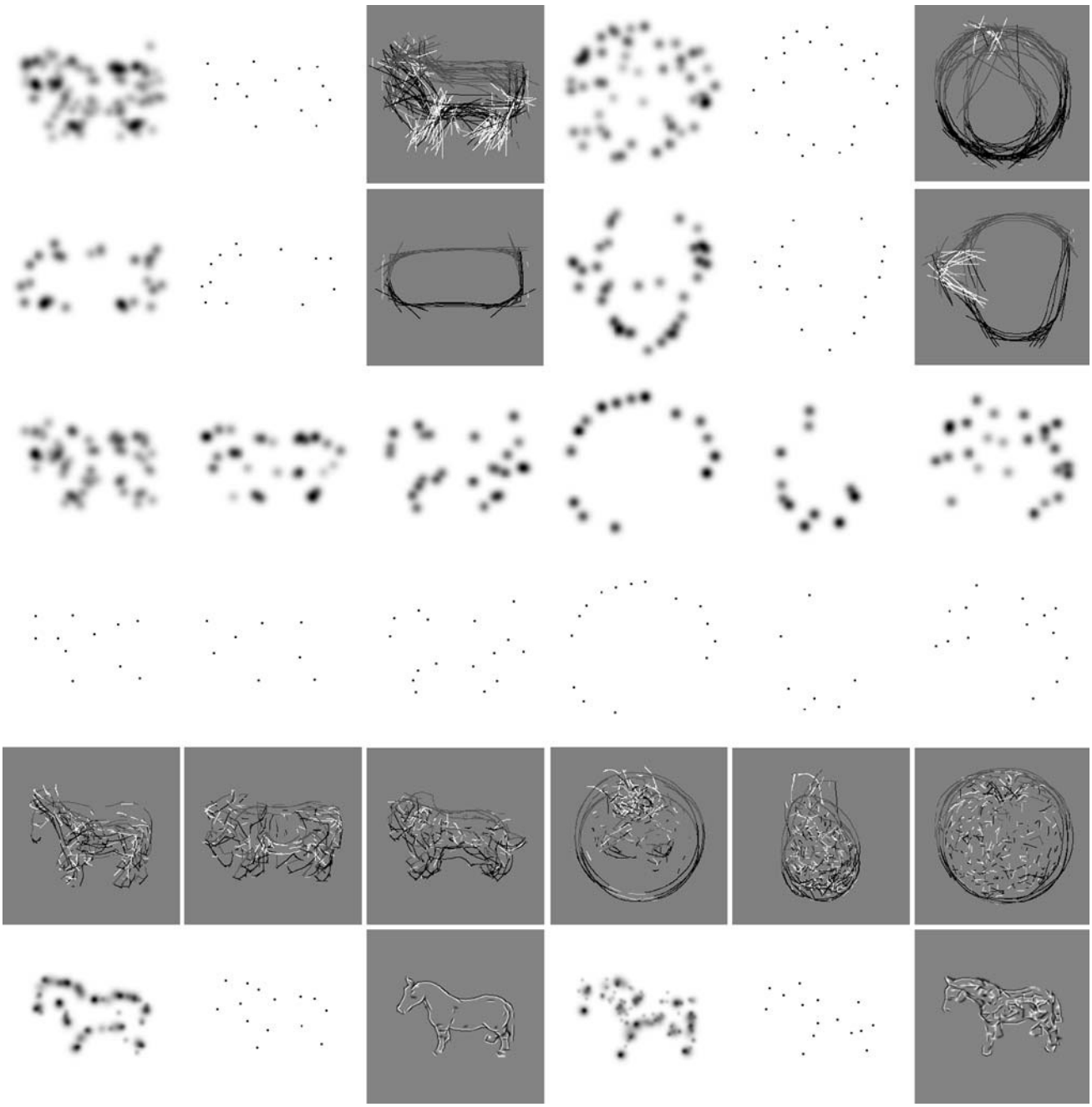

Fig. 5 Template data structure, showing only a single scale for each group. *Top two lines*: group templates for pre-categorization (animal, fruit, car and cup) at $\lambda = 32$. *Middle three lines*: the same for categorization (horse, cow, dog, tomato, pear and apple) at $\lambda = 8$. *Bottom line*: templates for final recognition, examples of two different horses at $\lambda = 4$

match (NO), its output will be blocked (X) and cannot reach the MAX block; this is done to prevent the system from selecting some arbitrary template when no template can match. Blocks marked "thr" perform thresholding, with four options: a very low value ($\ll$) implies the creation of a new template; a very high value ($\gg$) means final object recognition; if the value is not very much lower than the threshold ($<$), which means that more information is required to select the correct template, a feedback loop is activated (to the rightmost column of blocks, via FoA, in order to select more line/edge scales); if the value is not very much higher than the threshold ($>$), a specific template has been selected and this (group) template activates (selects) related (group) templates at the next level.

The heptagonal symbols between the LE and SM blocks of all templates represent comparisons (local and global correlations or matchings) between input and template features: line/edge events (LE) at categorization levels or

Journal : Large 10339 Dispatch : 18-5-2009 Pages : 19
Article No. : 262 ☐ LE ☐ TYPESET
MS Code : COGPRO-D-07-00012 ☑ CP ☑ DISK

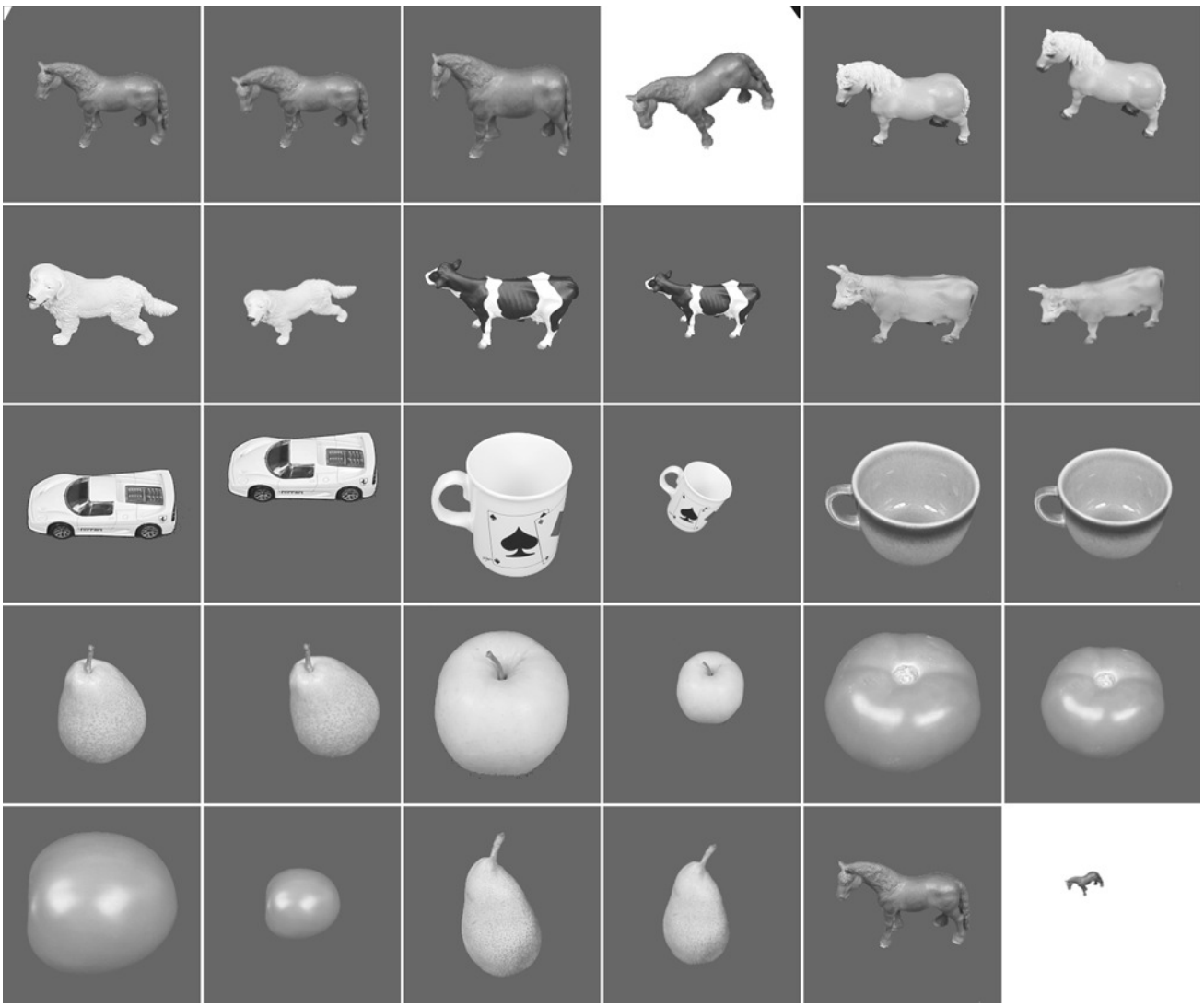

Fig. 6 Examples of objects used for categorization and recognition, with neighboring left-right columns showing the normalized and examples of modified objects

their symbolic representations (LErepr.) at the final recognition level. A comparison is only activated when a global match occurs, and after the dynamic routing of events as explained before. In the rightmost column of blocks, the following abbreviations are used: LF refers to the coarsest scales, AllF—to many scales (coarse, medium and fine) but in octave intervals, AllF to more scales with sub-octave intervals, and AllF + to the maximum number of scales with the smallest intervals. Instead of using only four selections, the number of scales is dynamic, i.e., more scales will be selected and used until the information provided by new scales becomes redundant.

With respect to visual pathways, the "where" path is more related to the detection, segregation, FoA and object-representation blocks in the rightmost column in Fig. 7, whereas the "what" path consists not only of the other blocks, but also the object-representation block. With respect to cortical areas involved, a strict attribution of the functional blocks to areas is still speculative, but a likely attribution is the following: simple, complex and end-stopped cells are located in area V1 (Olshausen and Field (2005). Line, edge and keypoint extractions also occur in V1, and possibly also in V2. More complex object representations, at least of important objects like faces, are established in V4, Chelazzi et al. (2001), and in IT, Zoccolan et al. (2005). FoA processing may start at the LGN level (before the cortex!) but is most pronounced in V4 and beyond Chelazzi et al. (2001), and figure-ground segregation may be achieved in V2, at least at the level of local occlusions; Qiu and von der Heydt (2005). Saliency maps may be present in MT, Born and Bradley (2005), and in PP, Deco and Rolls (2004), and global matching using templates in IT. Templates of groups

Journal : Large 10339 Dispatch : 18-5-2009 Pages : 19
Article No. : 262 LE TYPESET
MS Code : COGPRO-D-07-00012 CP DISK

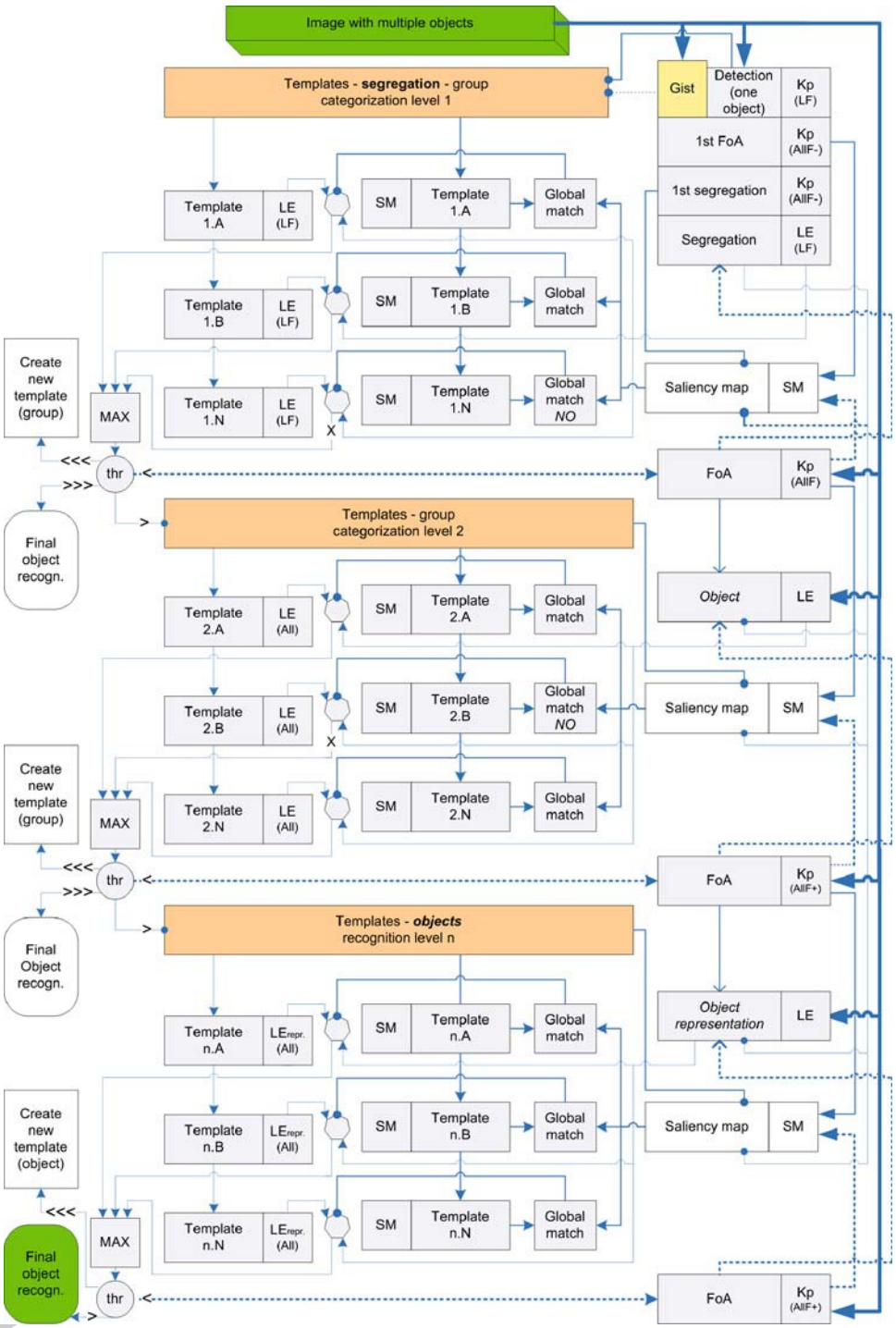

Fig. 7 Generalized architecture: blocks, features and information flow (see text)

and objects are stored—or at least available—at PF46 (orbitofrontal cortex); Miller (2000).

Discussion

There are many properties of a real-world scene that can be defined independently of the objects. For instance, a forest scene with trees can be described in terms of the degree of roughness and homogeneity of its textural components. Oliva and Torralba (2006) conclude that there is converging evidence that natural scene recognition may not depend on recognizing objects, and that the gist does not need to be built on top of the processing of individual objects.

Nevertheless, these processes are complementary. The initial gist can be the key for selecting the first group

Journal : Large 10339 Dispatch : 18-5-2009 Pages : 19
Article No. : 262 LE TYPESET
MS Code : COGPRO-D-07-00012 CP DISK

templates to start object recognition, but at some stage the objects should corroborate for the interpretation of the scene, and those objects must somehow be segregated. Any computational model of the cortical architecture should start with a model for getting the gist (forest scene), after which object recognition follows using segregated items, from generic information (trees) to more detailed information (tree type, leaf type). Only at the end of the entire process it may be possible to specify the gist; for example, a Mediterranean forest with tall pine trees.

Gist has not yet been implemented in our architecture, because we think that segregation of complex environments like natural scenes and gist are well interconnected processes. These processes may be based on complementary information channels which address motion and disparity, but also surface properties instead of structural object shape: (a) color processing in the cytochrome oxidase blobs, which are embedded in the cortical hypercolumns with simple, complex and end-stopped cells for line, edge and keypoint coding, must attribute colors to homogeneous (line/edge-free but also textured) object surfaces, and (b) texture coding based on specific groupings of outputs of grating cells in the case of rather periodic patterns, or other but similar processes in the case of more stochastic patterns. As shown by du Buf (2006), groupings of outputs of grating cells is a straightforward, data-driven and, therefore, fast bottom-up process which provides a segmentation (segregation) of linear, rectangular, and hexagonal textures. Therefore, a gist model, when seeing an image with blue and some white above green with a rather irregular pattern, may classify the scene, after sufficient training of course, as Mediterranean outdoor, thereby pre-selecting tree templates with a bias toward different pine trees (tall and more round etc.).

Not yet having a gist model, we simply assumed in our experiments that all group templates are available at the first categorization level, and that input objects are always seen against a homogeneous background (i.e. already segregated). At an early stage, only very coarse scales with big intervals are available, then medium scales with smaller intervals appear and finally the fine scales. The appearance and therefore the use of scales is directly related to all steps of the recognition process. The initial segregation starts with coarse scales, which provide a very diffuse object representation. This first segregation triggers a first categorization. When medium scales appear, and then fine scales, the segregation is improved and so is the categorization. The same occurs with the construction of the saliency map, first using keypoints detected at coarse scales and improving the map by adding keypoints detected at increasingly finer scales.

Invariance by neural routing from V1 via V2 to V4 etc. is based on the recurrent network layers used in the Deco and Rolls (2004) model, however, with one big difference: instead of only using simple cells (Gabor model) we apply explicit feature extractions and can use specific features to guide the routing. As a matter of fact, the routing can be seen as two vessels (input object and template) throwing anchors toward each other: the first, big anchor is the central object keypoint at very coarse scales and this is used to "position" the normalized template above the (shifted) input object. The second anchor is the most significant peak of the saliency map, obtained by summing keypoints over many scales, and this is used to match rotation and size. Once "anchored together," the "ropes" are used to steer many more ropes that connect specific structures of the vessels, like bow, rail and stern, in order to check whether the structures are similar and the vessels are of the same type.

Our "anchoring" method is similar to the theory developed by Olshausen et al. (1993), suggesting that the position and size of the reference frame can be set by the position and size of the object in the scene, assuming that the scene is at least roughly segmented, and that the orientation of the reference frame can be estimated from relatively low-level cues. The computational advantage of such a system is obvious: only a few views of an object need to be stored for recognition under different viewing conditions. The disadvantage, of course, is that a scene containing multiple objects requires serial processing, the system only being able to attend one object at a time. The same happens in our model and that of Deco and Rolls: dynamic routing steers the information flow by adapting neural interconnections in V2, etc. for some time, until recognition has been achieved, after which the adapted steering can be released for the inspection of another object (or region around a fixation point). Psychophysical evidence suggests that the brain, indeed, employs such a sequential strategy; Rensink (2000).

An interesting aspect of models is which features—and therefore which image representation—are being used. In our own model, explicit features are used: lines, edges, and keypoints are detected on the basis of responses of simple, complex, and end-stopped cells. The existence of other cells with very specific functions, like bar and grating cells, points at explicit feature extractions with increasing complexity at higher cortical areas; Rodrigues and du Buf (2006); du Buf (2006). The same idea, extended with increasing receptive field sizes, is supported by Deco and Rolls (2004), however, without explicit feature extractions. By only using simple cells (Gabor model), higher features are represented implicitly: complex cells group outputs of simple cells and end-stopped cells group outputs of complex cells. Nevertheless, in principle—they did not test this—their model should also be able to achieve invariant object recognition by combining feedback effects of top-

Journal : Large 10339 Dispatch : 18-5-2009 Pages : 19
Article No. : 262 LE TYPESET
MS Code : COGPRO-D-07-00012 CP DISK

down attentional mechanisms in a hierarchically organized set of cortical areas with convergent forward connectivity, reciprocal feedback connections, and local intra-area competition. As a consequence, we may say that these two models are converging, but eventually the same will happen with other computational models; Olshausen et al. (1993); Hamker (2005).

The templates used to illustrate the architecture were built from a very small database of 80 different objects. In future work the system must be extended and tested against a huge number of objects with many more categorization levels, thereby simulating a real application which approaches the challenge that our visual system faces every day. In this case, instead of having only three levels to obtain object recognition (pre-categorization, categorization and recognition), we will have n levels as shown in Fig. 7, each level only having N elements. An input object must not be compared with all the templates in the entire database, nor with a significant number of group templates. Part of the solution, not even mentioned until here, is to apply biasing of associative memory over time, as occurs in our brains: data streams in the case of frequently and recently seen objects are short and fast, whereas those in the case of occasionally and sparsely seen objects are longer and slower.

Every time that an object is recognized, its features could be added to all the matching templates. This way the system will be able to learn by updating the database using the most recent views of common objects. As discussed before, this can be done by biasing associative memory, but the memory itself must also be changed, for example by a weighted summation of new and old features, which can be fast in short-term memory but much slower in long-term memory. These ideas raise some problems which are not yet addressed by the present architecture: (a) for avoiding overgeneralization of the groups (classes), a threshold has to be implemented such that, before a class becomes too generic, it can be split (shown in Fig. 7 by "$\lll$"), and yet the two classes which provoked that split must still be generic enough for either class. (b) Objects with noise or occasional variations, like changing illumination or the deformation of non-rigid objects, or which are at the "edge" between two classes, are the ones which pose most problems. Therefore, their features should have a much smaller weight when contributing to the templates. (c) In general, different classes can have different weight factors, for example as a function of the number of objects that has been recognized within the class (group), creating the idea of a priority-secondary (but in reality continuous) organization, especially if temporal modulation (frequently and recently vs. occasionally and sparsely seen objects) is also applied. In addition, temporal changes, not only a sudden change of context, can lead to an evolution of the hierarchy of templates: they can shift forward toward early (coarse) categorization or back toward late (fine) categorization. (d) Finally, the contributions of all features can be weighted, especially when additional features like disparity (3D shape), texture and color will be included in the system. Apples vary between green, yellow, orange, red, and brown, so only blue can be excluded, but all oranges are orange and tomatoes are either green or red. However, such rules are not fixed because yellow and orange bell peppers appeared next to red and green ones only a few years ago in the supermarkets.

Summarizing, we presented a new model for invariant object categorization and recognition based on realistic multi-scale features which are extracted in the primary visual cortex. The model employs dynamic routing of features through the different layers to obtain 2D translation, rotation, and size invariance. The model was tested in the framework of an integrated and biologically plausible architecture in which information at coarse scales is used first and information at progressively finer scales later. By employing feedback loops, which are known to exist in abundance in the visual cortex, attention information based on keypoints and saliency maps is used to control the process. The entire process is composed of different categorization levels, recognition being the last one, with sequentially (but overlapping) coarse-to-fine-scale processing. Although not yet yielding perfect results, the architecture can deal with reasonable translations, rotations, and scalings. In a next step, the maximally allowable transformations must be determined, which depend on the number of neural layers used in the routing, and this will provide information on how many views of objects must be stored in memory.

Acknowledgments This research was supported by the Portuguese Foundation for Science and Technology (FCT), through the pluriannual funding of the Institute for Systems and Robotics (ISR/IST) by the POS_Conhecimento Program which includes FEDER funds, and by the FCT project PTDC/EIA/73633/2006 - SmartVision: active vision for the blind. We thank the anonymous reviewers for their useful comments.

Appendix Mathematical formulation of the model

Cell models and multi-scale feature extraction of Appendix

Gabor quadrature filters provide a model of cortical simple cells, Heitger et al. (1992). In the spatial domain the receptive field (RF) is denoted by (see also Rodrigues and du Buf (2006))

Journal : Large 10339 Dispatch : 18-5-2009 Pages : 19
Article No. : 262 LE TYPESET
MS Code : COGPRO-D-07-00012 CP DISK

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

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. We can apply a linear scaling between $f_{\min}$ and $f_{\max}$ with hundreds of contiguous scales. The scale of analysis is given in terms of λ expressed in pixels ($\lambda = 1$ corresponds to 1 pixel, and images presented have a size of 256 × 256 pixels).

Responses of even $R^E_{s,i}(x, y)$ (with $\phi = 0$) and odd $R^O_{s,i}(x, y)$ (with $\phi = -\pi/2$) simple cells are obtained by convolving the input image with luminance distribution $f(x, y)$ with the RFs,

$$\begin{aligned} R_{\lambda,\sigma,\theta,\phi}(x,y) &= f(x,y) \times g_{\lambda,\sigma,\theta,\phi}(x,y) \\ &= \iint_{\Omega} f(u,v) g_{\lambda,\sigma,\theta,\phi}(x-u, y-v)\mathrm{d}u\mathrm{d}v, \end{aligned} \quad (2)$$

s being the scale number were $s = 1$ corresponds to $\lambda \approx 4$, i the orientation ($\theta_i = i\pi/N_\theta$) and N_θ the number of orientations (here 8). Responses of complex cells are modeled by

$$C_{s,i}(x,y) = \sqrt{\left(R^E_{s,i}(x,y)\right)^2 + \left(R^O_{s,i}(x,y)\right)^2}. \quad (3)$$

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

$$\begin{aligned} &S_{s,i}(x,y) \\ &= \left[C_{s,i}\left(x + \mathrm{d}\hat{S}_{s,i}, y - \mathrm{d}\hat{C}_{s,i}\right) - C_{s,i}\left(x - \mathrm{d}\hat{S}_{s,i}, y + \mathrm{d}\hat{C}_{s,i}\right)\right]^+, \end{aligned} \quad (4)$$

and

$$\begin{aligned} D_{s,i}(x,y) = \Bigg[& C_{s,i}(x,y) - \frac{1}{2}C_{s,i}\left(x + 2\mathrm{d}\hat{S}_{s,i}, y - 2\mathrm{d}\hat{C}_{s,i}\right) \\ & -\frac{1}{2}C_{s,i}\left(x - 2d\hat{S}_{s,i}, y + 2d\hat{C}_{s,i}\right)\Bigg]^+. \end{aligned} \quad (5)$$

The responses of the two cell types are obtained by $S_s(x,y) = \sum_{i=0}^{N_\theta-1} S_{s,i}(x,y)$ and $D_s(x,y) = \sum_{i=0}^{N_\theta-1} D_{s,i}(x,y)$. The distance d is scaled linearly with the filter scale s (we use $d = 0.6$ s). To compute the keypoint maps, all responses of end-stopped cells along straight lines and edges are suppressed, for which tangential (T) and radial (R) inhibitions are used:

$$\begin{aligned} I^T_s(x,y) = \sum_{i=0}^{2N_\theta-1} \Bigg[& -C_{s,i\mathrm{mod}N_\theta}(x,y) \\ & + C_{s,i\mathrm{mod}N_\theta}\left(x + d\hat{C}_{s,i}, y + d\hat{S}_{s,i}\right)\Bigg]^+, \end{aligned} \quad (6)$$

and

$$\begin{aligned} I^R_s(x,y) = \sum_{i=0}^{2N_\theta-1} \Bigg[& C_{s,i\mathrm{mod}N_\theta}(x,y) - 4C_{s,(i+N_\theta/2)\mathrm{mod}N_\theta} \\ & \times \left(x + \frac{d}{2}\hat{C}_{s,i}, y + \frac{d}{2}\hat{S}_{s,i}\right)\Bigg]^+. \end{aligned} \quad (7)$$

Then we apply $I_s = I^T_s + I^R_s$ for obtaining the end-stopped maps $K^S_s(x, y) = S_s(x, y) - gI_s(x, y)$ and $K^D_s(x, y) = D_s(x, y) - gI_s(x, y)$, with $g \approx 1.0$, and the combined map $K^R_s(x, y) = \max\{K^S_s(x, y), K^D_s(x, y)\}$. In the last step, local maxima of $K^R_s(x, y)$ in x and y are detected to obtain each single point (marked white in Fig. 1, 3rd row) which represents a keypoint at each scale s.

The line/edge maps are obtained on the basis of the responses $R^E_{s,i_d}(x,y)$ and $R^O_{s,i_d}(x,y)$, were i_d is the dominant orientation of $\breve{C}_{s,i} = \left[C_{s,i}(x,y) - \beta\left(I^L_{s,i}(x,y) + I^E_{s,i}(x,y)\right)\right]^+$, i.e., the orientation with the maximum response of $\breve{C}_{s,i}$, where

$$\begin{aligned} I^L_{s,i}(x,y) = & \left[C_{s,i}\left(x + d\hat{C}_{s,i}, y + d\hat{S}_{s,i}\right)\right. \\ & \left. - C_{s,i}\left(x - d\hat{C}_{s,i}, y - d\hat{S}_{s,i}\right)\right]^+ \\ & + \left[C_{s,i}\left(x - d\hat{C}_{s,i}, y - d\hat{S}_{s,i}\right)\right. \\ & \left. - C_{s,i}\left(x + d\hat{C}_{s,i}, y + d\hat{S}_{s,i}\right)\right]^+ \end{aligned} \quad (8)$$

and

$$\begin{aligned} I^C_{s,i}(x,y) = & \left[C_{s,i\mathrm{mod}N_\theta}\left(x + 2d\hat{C}_{s,i}, y + 2d\hat{S}_{s,i}\right)\right. \\ & \left. - 2C_{s,i}(x,y) + C_{s,i\mathrm{mod}N_\theta}\left(x - 2d\hat{C}_{s,i}, y - 2d\hat{S}_{s,i}\right)\right]^+ \end{aligned} \quad (9)$$

denote lateral (L) and cross-orientation (C) inhibition, which are necessary because simple and complex cells respond beyond line and edge terminations, for example beyond the corners of a rectangle, see Rodrigues and du Buf (2008), using $\beta \approx 1$. At each position (x, y) for which $\breve{C}_s > 0$, with

$$\breve{C}_s = \sum_{i=0}^{N_\theta-1} \breve{C}_{s,i}, \quad (10)$$

the event type and polarity are determined by checking the responses of the simple cells $R^E_{s,i_d}(x,y)$ and $R^O_{s,i_d}(x,y)$ for a local maximum (or minimum by rectification) using a dendritic field size of $\pm\lambda/4$. Exactly the same condition

Journal : Large 10339 Dispatch : 18-5-2009 Pages : 19
Article No. : 262 LE TYPESET
MS Code : COGPRO-D-07-00012 CP DISK

(maximum) on the basis of responses of complex cells ($\breve{C}_s$) has to be checked. Finally a coinciding zero-crossing (z.c.) in $R^O_{s,i_d}(x,y)$ or $R^E_{s,i_d}(x,y)$, again on $\pm\lambda/4$, must occur. Summarizing, four event types can be detected: positive line $L_{s,+}(\breve{C}_s = \text{Max};\ R^O_{s,i_d} = \text{z.c.};\ R^E_{s,i_d} = \text{Max})$, negative line $L_{s,-}(\breve{C}_s = \text{Max};\ R^O_{s,i_d} = \text{z.c.};\ R^E_{s,i_d} = \text{Min})$, positive edge $E_{s,+}(\breve{C}_s = \text{Max};\ R^E_{s,i_d} = \text{z.c.};\ R^O_{s,i_d} = \text{Max})$, and negative edge $E_{s,-}(\breve{C}_s = \text{Max};\ R^E_{s,i_d} = \text{z.c.};\ R^O_{s,i_d} = \text{Min})$.

Reconstruction model of Appendix

As explained in the Section "Lines, edges, keypoints and saliency maps", we can assume that the visual system extracts lines and edges for object recognition, and that responding "line cells" and "edge cells" are also interpreted symbolically for creating a brightness representation; see du Buf and Fisher (1995) and Rodrigues and du Buf (2008) for more details. The 2D line and edge representations (positive; negative ones are obtained by multiplication by -1) were implemented on the basis of 1D cross-profiles. Using the normal definition of a Gaussian in x,

$$G(x;\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-x^2}{2\sigma^2}\right), \tag{11}$$

a generalized positive line (in 1D) is described by $\Psi_s(x) = G(x;\ s\sigma_l)$ where σ_l defines the width of the line profile. Similarly, a generalized positive edge with width σ_e is defined by $\Lambda_s(x) = G(x; s\sigma_e)\cdot\Phi\left(x/s\sigma_e\sqrt{2}\right)$, where $\Phi(z)$ is the (generally complex) error function

$$\Phi(z) = \frac{2}{\sqrt{\pi}}\int_0^z \mathrm{e}^{-t^2}\mathrm{d}t. \tag{12}$$

For each of the four event maps $L_{s,\pm}(x,y)$ and $E_{s,\pm}(x,y)$, the corresponding 1D profiles $\Psi_{s,\pm}(x)$ and $\Lambda_{s,\pm}(x)$ are rotated to the dominant orientation i_d, and multiplied by the amplitude of the complex cells $C_{s,i_d}(x,y)$ at the detected positions. For generating the 2D representation maps $\Psi_{s,\pm}(x,\ y)$ and $\Lambda_{s,\pm}(x,\ y)$, it is necessary to interpolate values between two consecutive profiles (of neighboring cells) such that gaps are filled. In Fig. 1, the third and fourth image from left on the second row show the summation of the representations at a fine and a coarse scale. Final image reconstruction R is obtained by

$$\begin{aligned} R(x,y) = {} & \gamma\ \mathrm{LP}(x,y;\sigma_r) \\ & + \frac{(1-\gamma)}{N_s}\sum_{s=1}^{N_s}\left[\left(\Psi_{s,+}(x,y) + \Psi_{s,-}(x,y)\right.\right. \\ & \left.\left. + \Lambda_{s,+}(x,y) + \Lambda_{s,-}(x,y)\right)\right] \end{aligned} \tag{13}$$

with $\mathrm{LP}(x,\ y;\ \sigma_r)$ a Gaussian-filtered lowpass image, γ a coefficient which balances the lowpass component and the line/edge representations, and N_s the number of scales used. The rightmost image on the middle row in Fig. 1 was obtained with $\gamma = 0.5$, $\sigma_r = 5$ and $N_s = 8$ ($\lambda = \{4;\ 8;\ 12;\ 16;\ 20;\ 24;\ 28;\ 32\}$). Processes like the 2D interpolation of 1D cross-profiles are speculative, but they are necessary for showing 2D images; for more details see Rodrigues and du Buf (2008).

Focus-of-Attention by saliency maps of Appendix

For modeling Focus-of-Attention (FoA) we need a map, called saliency map, which indicates the most important points to be analyzed (fixated); Rodrigues and du Buf (2006). Activities of all keypoint cells at position $(x,\ y)$ can be summed over scales s by grouping cells. Each keypoint has a Region-of-Interest (RoI) that can be used to process—during an eye fixation—other information inside the RoI. The RoI is small at fine scales and big at coarse scales. This is modeled by assuming circular axonal fields of keypoint cells, with a size of 3×3 at the finest scale $\lambda = 4$, but simulated by using a 2D Gaussian with σ_{sm} and with linear scaling toward coarser scales. The saliency map SM is obtained by

$$\mathrm{SM}(x,y) = \sum_{s=1}^{N_s} K_s(x,y) \times G(x,y;s\sigma_{\mathrm{sm}}). \tag{14}$$

Template data structure of Appendix

The template data structure T depends on the type of process (ty), i.e., pre-categorization (PC), categorization (C) and recognition (R), and consequently also on the group of objects used, Ω^{ty}_k, with k the group number. In addition, it has the following elements: (a) the peaks of the saliency map (PSM), (b) the central keypoint (CKP) and (c) the line and edge information.

Prior to computing the peak (PSM) information, the saliency map for each entire group was obtained by summing the SMs of all the elements from the group

$$T^{\mathrm{ty}}_{\mathrm{SM},k}(x,y) = \sum_{n\in\Omega^{\mathrm{ty}}_k}\mathrm{SM}_n(x,y). \tag{15}$$

Each peak $T^{\mathrm{ty}}_{\mathrm{PSM},k}(x,\ y)$ of the SM is a single point (see e.g. Fig. 4b, e, g and i), i.e., a local maximum in x and y of the $T^{\mathrm{ty}}_{\mathrm{SM},k}$ map with non-maximum suppression and thresholding. The central keypoint corresponds to a location close to an object's centroid $(x_{c_{\mathrm{ty},k}}, y_{c_{\mathrm{ty},k}})$. This is different for each group and for each level (ty) of processing. The scale of the CKP is determined when only a single keypoint exists in the entire image, which only occurs at a very coarse scale, see Rodrigues and du Buf (2006), and this scale may be different for each group k and level of recognition ty.

Journal : Large 10339 | Dispatch : 18-5-2009 | Pages : 19
Article No. : 262 | LE | TYPESET
MS Code : COGPRO-D-07-00012 | CP | DISK

Line/edge (LE) templates for each group also depend on the processing level ty. For pre-categorization, detected lines (L^{PC}) and edges (E^{PC}) were considered as binary events (no polarity or event type was applied) from the segregated (segr.) but normalized objects of the corresponding group, with the scales s_i corresponding to $\lambda = \{24; 28; 32\}$. For categorization, L^C and E^C were also considered as binary events (no polarity and event type), but now from the original objects. Finally, for L^R and E^R we used the corresponding LE representations $\Psi_{\pm}$ and $\Lambda_{\pm}$ (with polarity and event type). In the last two cases s_i corresponded to all the scales (see Table 1, rows 7–9). Line and edge information can be formalized as

$$T^{ty}_{LE,s_i,k}(x,y) = \sum_{n\in\Omega^{ty}_k}\Big[L^{ty}_{s_i,+,n}(x,y) + E^{ty}_{s_i,+,n}(x,y) + L^{ty}_{s_i,-,n}(x,y) + E^{ty}_{s_i,-,n}(x,y)\Big]. \quad (16)$$

The final template stored in memory can be summarized as

$$T(x,y) = \bigcup_{k=1}^{4}\left\{T^{PC}_{PSM,k}(x,y), T^{PC}_{CKP,k}(x_{c_{PC,k}}, y_{c_{PC,k}}), T^{PC}_{LE,s_i,k}(x,y)\right\} \bigcup_{k=1}^{8}\left\{T^{C}_{PSM,k}(x,y), T^{C}_{CKP,k}(x_{c_{C,k}}, y_{c_{C,k}}), T^{C}_{LE,s_i,k}(x,y)\right\} \bigcup_{k=1}^{80}\left\{T^{R}_{PSM,k}(x,y), T^{R}_{CKP,k}(x_{c_{R,k}}, y_{c_{R,k}}), T^{R}_{LE,s_i,k}(x,y)\right\}. \quad (17)$$

The numbers 4, 8 and 80 in Eq. 20 correspond to the number of groups at each recognition level (ty) of the present database.

Dynamic routing of Appendix

The dynamic routing process is explained in detail in Section "Invariant object categorization and recognition". We can therefore skip the normal 2D translation, rotation and size transformations, and only explain the final step. This step serves to check whether there exist significant pairs of coinciding peaks between object and template, for which two thresholds are applied: (a) all object peaks O^{ty}_{PSM} with amplitude less than half of the maximum amplitude of the template peaks $T^{ty}_{PSM,k}$ are inhibited. (b) Remaining peaks of object and templates are checked whether pairs coincide in a Gaussian window, the size (ω) of which corresponds to the envelope of simple and complex cells at the middle scale used for each level ty of recognition:

$$\vartheta^{ty}_k = \sum\left(O^{ty}_{PSM}\bigcap_{\omega}T^{ty}_{PSM,k}\right). \quad (18)$$

Table 1 Top 6 rows: Group templates, for pre-categorization (PC), categorization (*C*) and recognition (*R*)

Ω^{PC}_k	15 FR_{segr} ($k = 1$)								5 CU_{segr} ($k = 2$)		
Ω^{C}_k	5 AP ($k = 1$)		5 PE ($k = 2$)			5 TO ($k = 3$)			5 CU ($k = 4$)		
Ω^{R}_k	AP1 ($k = 1$)	… ($k = 2,\ldots,9$)	AP10 ($k = 10$)	PE1 ($k = 11$)	… ($k = 12,\ldots,19$)	PE10 ($k = 20$)	TO1 ($k = 21$)	… ($k = 22,\ldots,29$)	TO10 ($k = 30$)	CU1 ($k = 31$)	… ($k = 31,\ldots,39$)	CU10 ($k = 40$)
Ω^{PC}_k	15 AN_{segr} ($k = 3$)									5 CA_{segr} ($k = 4$)		
Ω^{C}_k	5 DO ($k = 5$)			5 CO ($k = 7$)			5 HO ($k = 6$)			5 CA ($k = 8$)		
Ω^{R}_k	DO1 ($k = 41$)	… ($k = 42,\ldots,49$)	DO10 ($k = 50$)	CO1 ($k = 51$)	… ($k = 52,\ldots,59$)	CO10 ($k = 60$)	HO1 ($k = 61$)	… ($k = 62,\ldots,69$)	HO10 ($k = 70$)	CA1 ($k = 71$)	… ($k = 72,\ldots,79$)	CA10 ($k = 80$)
L^{PC};E^{PC}; s_i	$L_{segr.,s}$; $E_{segr.,s}$; $\pm$; $\lambda = \{24; 28; 32\}$											
L^{C};E^{C};s_i	$L_{s,\pm}$; $E_{s,\pm}$; $\lambda = \{4; 8; 12; 16; 20; 24; 28; 32\}$											
L^{R};E^{R};s_i	$\Psi_{s,\pm}$; $\Lambda_{s,\pm}$; $\lambda = \{4; 8; 12; 16; 20; 24; 28; 32\}$											

The next 3 lines specify line/edge types and scales used to create the templates

AN animal, *DO* dog, *CO* cow, *HO* horse, *FR* fruit, *AP* apple, *PE* pear, *TO* tomato, *CA* car, *CU* cup

Journal : Large 10339 Dispatch : 18-5-2009 Pages : 19
Article No. : 262 LE TYPESET
MS Code : COGPRO-D-07-00012 CP DISK

If ϑ_k^{ty} is higher then half of the sum of all SM peaks in $T_{\text{PSM},k}^{\text{ty}}$ which passed the first threshold, i.e., $\vartheta_k^{\text{ty}} > 1/2 \sum T_{\text{PSM},k}^{\text{ty}}$, a possible match can occur, and the next step is to determine the similarity between object and templates (next section) for each of the different k and ty. If the maximum similarity is not enough, object and template may be different, but first the maximum peak of the object is inhibited and a new maximum is used to test other possible combinations.

Similarity between objects and templates of Appendix

The similarity between an un-normalized object and a normalized template is determined every time that dynamic routing has been established, i.e. the similarity is computed for all the k's where a possible match occurs at the same level ty. This means that the line/edge (LE) information of the object at all scales is translated, scaled and rotated (TSR), as explained in Section "Invariant object categorization and recognition" for the SM peaks. We denote this LE transformation by $\text{OL}_{\text{TSR},s,\pm}^{\text{ty}}$ and $\text{OE}_{\text{TSR},s,\pm}^{\text{ty}}$, where O stands for object (or T for template), L for lines and E for edges, which can be positive (+) and negative (-).

For each group template, at each of the scales, a positional relaxation area (RT) was created around each responding event cell, by assuming grouping cells with a dendritic field size, again modeled by a 2D Gaussian function, coupled to the size of the underlying complex cells:

$$\text{RT}_{\text{LE},s_i,k}^{(\text{ty}\wedge\text{ty}\neq R)}(x,y) = T_{\text{LE},s_i,k}^{\text{ty}}(x,y) \times G(x,y;s_i\sigma). \tag{19}$$

These grouping cells sum the occurrence of object events around template events, which can be seen as a local correlation, and then activities of all grouping cells are summed to obtain the global correlation

$$\Gamma_{s_i,k}^{\text{ty}} = \sum_{x,y} \left[\text{RT}_{\text{LE},s_i,k}^{\text{ty}}(x,y) \bigcap_{x,y} \left[\text{OL}_{\text{TSR},s,\pm}^{\text{ty}}(x,y) + \text{OE}_{\text{TSR},s,\pm}(x,y) \right] \right]. \tag{20}$$

These final groupings are compared over the k templates, scale by scale, and the template with maximum response is selected, $K_{s_i}^{\text{ty}} = \max_k \left\{ \Gamma_{s_i,k}^{\text{ty}} \right\}$. Finally, the template with the maximum number of correspondences over the scales s_i is selected, $K^{ty} = \max_{s_i} \left\{ \sum K_{s_i}^{\text{ty}} \right\}$.

This K^{ty} is a single number, i.e. the group number of the corresponding template: 1–4 for ty = PC, 1–8 for ty = C and 1–80 for ty = R.

It should be stressed that for recognition the process is the same, except that the relaxation area in Eq. 19 is applied to line and edge cells (see Section "Reconstruction model"), and in Eq. 20 the events must have the same position, type,and polarity, i.e., $\bigcap_{x,y;\pm;L/E}$.

References

Bar M (2003) A cortical mechanism for triggering top-down facilitation in visual object recognition. J Cogn Neurosci 15(4):600–609. doi:10.1162/089892903321662976

Bar M (2004) Visual objects in context. Nat Rev Neurosci 5:619–629. doi:10.1038/nrn1476

Bar M, Kassam K, Ghuman A, Boshyan J, Schmid A, Dale A, Hämäläinen M, Marinkovic K, Schacter D, Rosen B, Halgren E (2006) Top-down facilitation of visual recognition. Proc Natl Acad Sci USA 103(2):449–454. doi:10.1073/pnas.0507062103

Berson D (2003) Strange vision: ganglion cells as circadian photoreceptors. Trends Neurosci 26(6):314–320. doi:10.1016/S0166-2236(03)00130-9

Born R, Bradley D (2005) Structure and function of visual area MT. Annu Rev Neurosci 28:157–189. doi:10.1146/annurev.neuro.26.041002.131052

Chelazzi L, Miller E, Duncan J, Desimone R (2001) Responses of neurons in macaque area V4 during memory-guided visual search. Cereb Cortex 11(8):761–772. doi:10.1093/cercor/11.8.761

Cox D, Meier P, Oertelt N, DiCarlo J (2005) 'Breaking' position-invariant object recognition. Nat Neurosci 8(9):1145–1147. doi:10.1038/nn1519

Deco G, Rolls E (2004) A neurodynamical cortical model of visual attention and invariant object recognition. Vision Res (44):621–642. doi:10.1016/j.visres.2003.09.037

Deco G, Rolls E (2005) Attention, short term memory, and action selection: a unifying theory. Prog Neurobiol 76:236–256

du Buf J (2006) Improved grating and bar cell models in cortical area V1 and texture coding. Image Vis Comput. doi:10.1016/j.imavis.2006.06.005

du Buf J, Fisher S (1995) Modeling brightness perception and syntactical image coding. Optical Eng 34(7):1900–1911. doi:10.1117/12.200602

Grill-Spector K, Kanwisher N (2005) Visual recognition: as soon as you know it is there, you know what it is. Psychol Sci 16(2):152–160. doi:10.1111/j.0956-7976.2005.00796.x

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. Cereb Cortex 15:431–447. doi:10.1093/cercor/bhh146

Heitger F, Rosenthaler L, von der Heydt R, Peterhans E, Kubler O (1992) Simulation of neural contour mechanisms: from simple to end-stopped cells. Vis Res 32(5):963–981. doi:10.1016/0042-6989(92)90039-L

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

Lindeberg T (1994) Scale-space theory in computer vision. Kluwer Academic Publishers, Dordrecht

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

Miller E (2000) The prefrontal cortex and cognitive control. Nat Rev Neurosci 1(1):59–65. doi:10.1038/35036228

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

Journal : Large 10339 | Dispatch : 18-5-2009 | Pages : 19
Article No. : 262 | LE | TYPESET
MS Code : COGPRO-D-07-00012 | CP | DISK

Olshausen B, Field D (2005) How close are we to understanding V1? Neural Comput 17(8):1665–1699. doi:10.1162/0899766054026639

Olshausen B, Anderson C, van Essen D (1993) A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. J Neurosci 13(11):4700–4719

Prime D, Ward L (2006) Cortical expressions of inhibition of return. Brain Res 1072(1):161–174. doi:10.1016/j.brainres.2005.11.081

Qiu FT, von der Heydt R (2005) Figure and ground in the visual cortex: V2 combines stereoscopic cues with gestalt rules. Neuron 47(1):155–166. doi:10.1016/j.neuron.2005.05.028

Rehn M, Sommer F (2006) Storing and restoring visual input with collaborative rank coding and associative memory. Neurocomputing 69(10–12):1219–1223. doi:10.1016/j.neucom.2005.12.080

Rensink R (2000) The dynamic representation of scenes. Vis Cogn 7(1–3):17–42. doi:10.1080/135062800394667

Riesenhuber M, Poggio T (2000) CBF: a new framework for object categorization in cortex. In: Proc. 1st IEEE international workshop, biologically motivated computer vision, Seoul, Korea, May 15–17, pp 1–9

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

Rodrigues J, du Buf J (2005) Multi-scale keypoints in V1 and face detection. In: Proc. 1st int. symp. brain, vision and artif. intell. Springer LNCS, vol 3704. Naples (Italy), pp 205–214

Rodrigues J, du Buf J (2006) 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

Rodrigues J, du Buf J (2008) Multi-scale lines and edges in V1 and beyond: brightness, object categorization and recognition, and consciousness. Biosystems (accepted). doi:10.1016/j.biosystems.2008.10.006

Serre T, Riesenhuber M (2004) Realistic modeling of simple and complex cell tuning in the Hmax model, and implications for invariant object recognition in cortex. CBCL Paper 239/AI Memo 2004-017, Massachusetts Institute of Technology, Cambridge

Stringer S, Perry G, Rolls E, Proske H (2006) Learning invariant object recognition in the visual system with continuous transformations. Biol Cybern 94:128–142. doi:10.1007/s00422-005-0030-z

Tarr J (2005) How experience shapes vision. Psychol Sci Agenda 19(7)

Walther D, Rutishauser U, Koch C, Perona P (2005) Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. Comput Vis Image Underst 100(1–2):41–63. doi:10.1016/j.cviu.2004.09.004

Zoccolan D, Cox D, DiCarlo J (2005) Multiple object response normalization in monkey inferotemporal cortex. J Neurosci 25(36):8150–8164. doi:10.1523/JNEUROSCI.2058-05.2005

Journal : Large 10339 | Dispatch : 18-5-2009 | Pages : 19
Article No. : 262 | ☐ LE | ☐ TYPESET
MS Code : COGPRO-D-07-00012 | ☑ CP | ☑ DISK