

# OBJECT REPRESENTATION-BY-FRAGMENTS IN THE VISUAL SYSTEM: A NEUROCOMPUTATIONAL MODEL

Dan W. Joyce, Lynn V. Richards, Angelo Cangelosi and Kenny R. Coventry

Centre for Neural and Adaptive Systems / Centre for Thinking and Language  
School of Computing, Dept. Of Psychology and Plymouth Institute of Neuroscience  
University of Plymouth, UK

## ABSTRACT

This paper presents a model of visual object representation by *fragment* views, rather than canonically-oriented whole-object views used in Chorus systems [1, 3]. Following recent results [2] on object representation in inferotemporal cells during free viewing, we implemented a simplified attentional system which yields *fragment* views of objects, which are then used to train object-tuned modules. Each object is represented by a complete RBF module cf. [1, 3], instantiating a representation space. We show that such a system can produce distributed representations, like Chorus of views systems, and that dissociating objects from retinotopy enables a fuller model of scene geometry analysis to be advanced.

## 1. INTRODUCTION

The Chorus model [1] proposes that object representation in the primate visual system can be simulated by RBF modules, trained to respond for views of objects. We use Edelman's theory and implementation strategy, but in a dynamic context. A Chorus system operating in an active visual system, performing online visual scene analysis with respect to objects and locations, must account for the fact that objects will not be centred in the visual input, and indeed, the whole object may not be available to the high resolution, ventral object identity stream. The kind of model we envisage is congruent (in terms of time scale and representational properties) with [2] which describes an experiment where monkeys are trained on an object set, and then allowed to free saccade around a visual scene. Intra-temporal (IT) neurons were shown to respond to objects as the monkey fixated. Our model simulates this free-viewing process, and trains on fragments (rather than whole views).

A tentative model of how the ventral and dorsal pathways might interact in order that a visual scene can be perceived was given in [4]. If a stable array of locations and

objects is to be aggregated into a visual scene "percept" (e.g. perceiving the teapot, teacup and liquid scene shown in Figure 2) then the ventral (what) stream has central visual field acuity (ideal for detailed object representation and recognition) but lacks the dorsal stream's parietal, heteromodal information concerning motor information (presumably with respect to reaching and saccadic eye movements) which aids to stabilise the spatial array. This idea was taken up in the closing dialogues of [1], and more recently, in [5]. Our model, we propose, is a sound basis for exploring this proposal.

## 2. EARLY VISUAL PROCESSING

The model processes sequences of images represented as  $800 \times 600$  pixels in 256 greyscale-levels which are extracted from videos showing real objects moving in prescribed ways, or showing interactions between a number of the objects. For example, one sequence used to train the model was a video showing a teapot pouring liquid into a teacup. Each object (the cup, teapot and liquid) were photographed separately, and then composed against a blank background using an animation production application, producing short videos of 5-6 seconds in duration. This enables us to produce videos which are suitable for the model (e.g. we are able to ignore complex problems such as object segmentation) and plausible to human observers. The videos are then presented to the model as sequences of images. The model uses 3 scales of a Laplacian of Gaussian filter. This results in a high-resolution, retinotopic array at 3 scales. Further reduction is obtained by sub-sampling all three scales to 25% of the original size, resulting in a sequence of images at 200 by 150 pixels, at 3 scales.

## 3. VISUAL PROCESSING

Form and motion processing are roughly hierarchically organised in primates [8, 9, 10, 11, 6, 7]. Recent findings with macaques also show that V2 cells in macaques appear to be selective for *complex* contour stimuli [12] and V3

---

This work was supported by EPSRC Grant Number GR/N38145. Corresponding author Dan Joyce, email: dan.j@soc.plymouth.ac.uk

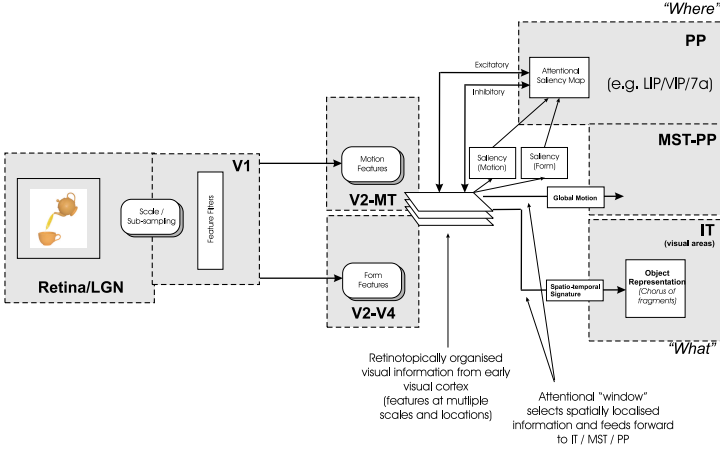


Figure 1: Model Schematic

showed sensitivity to higher temporal frequencies, and like MT showed preference for two dimensional movements, rather than for uni-dimensional motion in V1 [13].

Given that our model would use sequences of images, and following from recent psychophysical evidence about the role of spatio-temporal information for object recognition [14], we implemented a form and motion processing system as follows. Form filters produce *edge*, *region boundary*, and *texture* information and motion filters yield *motion magnitude* and *direction* information at each spatial location. For brevity, the precise implementations of these operators is given elsewhere [15]. Motion direction and motion magnitude are modelled by gradient-sensitive cells operating with first-derivative information [16]. Our aim is not to reconstruct the optical flow field *per se*, but to detect motion which is spatio-temporally localised (i.e. by hypothesizing a cell which is selective to a preferred direction, centred at some image location and sampling over some short time differential).

So, for any fixated region in a  $200 \times 150$  visual representation, the model has access to local motion, edge, texture and ‘boundary’ information (the latter being a representation where activations are uniform for a region of intensity, and zero otherwise; a kind of surface-minus-edge representation).

#### 4. A SIMPLIFIED ATTENTIONAL SYSTEM SUBSERVING A CHORUS-LIKE MODEL

The results of the above operations result in a potentially massive array of information to process. At each location in each of the  $200 \times 150$  sub-sampled images there will be 4 filter outputs (edges, texture, region and motion) totalling some 120,000 measurements per image in a sequence. We simulate something like the putative processing occurring in the free-viewing scenario [2].

We implemented a simple, stimulus-driven model of selective attention which we propose implements the kinds of covert attention associated with saccadic eye movements i.e. of temporal order 200-300ms. This is a simplified and quite heuristic version of specific models of visual attention, cf. [17], [18], and [10] which typically model either *overt* or *covert* attention. To proceed, we require some way of simulating fixation, selectively enhancing activations in the early visual representation of the fixated region, and further processing simulating the ventral processing of this information. We view this as a stimulus-driven, or exogenous, process. We define a saliency map which aids the directing of attention. In addition to the early retinotopic visual representations described above, a much coarser representation is maintained at approximately 10% of the original image resolution. There is evidence for a dorsal extra-striate pathway [19] from V1 to the medial and lateral parietal cortex to guide vision-for-action (including shifting attention by gaze). Similarly, [20] and [21] review the involvement of a dorsal fronto-parietal network of ventral intraparietal (VIP), lateral intra-parietal (LIP) and area 7a. [22] discusses a similar network, which integrates multi-modal representations of space, where LIP neurons code using a body-centred frame of reference, and area 7a (with parahippocampal connectivity) uses world-centred frame of reference. [23] similarly suggests that a faster, coarse representation might be used in the dorsal-parietal stream to direct attention for the slower but more spatially accurate ventral stream. We hypothesise that some representation (similar to the one implemented here) is available to direct attention (and distributed amongst such parietal networks). This separate, derivative representation is maintained for directing and shifting attention i.e. a representation for directing attention (in the dorsal-parietal stream) *versus* a representation subserving visual representation/recognition tasks (in the predominantly ventral, temporal pathway).

#### 4.1. Saliency and Fragment Extraction

Two saliency maps represent local image intensity and motion information at any time interval. These are linearly combined such that the brightest moving point is the most likely candidate for the next fixation. A winner-take-all system selects the best candidate from the combined saliency. Fixation is shifted to that location scaled into the co-ordinates of the higher resolution visual buffer. An inhibition system is simulated using a simple leaky-integrator model. At any image location, the inhibition signal prevents re-attending (in some time frame) by a Gaussian function of the distance to the last fixation (governed by a spread parameter  $\sigma_w$ ) and modulated by a decay parameter  $\lambda$  which controls how fast inhibition fades.

At the fixated location, the high resolution feature filters are compressed by Gaussian receptive fields into a  $5 \times 5$  rep-

resentation for each filter (edge, texture, motion etc.). This results in a 125 dimensional representation of a *fragment* of a fixated region. This is similar to Edelman’s mapping from image to measurement space, but the extraction of this representation is by the dynamic system described (rather than by ensuring each object is presented statically at the center of the display). The system is summarised in Figure 1.

Figure 2 shows a snapshot of the overall saliency map (with inhibition) after 9 fixations. Regions with inhibited saliency are shown as shades of grey, with highest salience shown as white. The line and numbers show the order of fixations as the video progressed. The video shows liquid pouring from a teapot (top right) into a cup (bottom left).

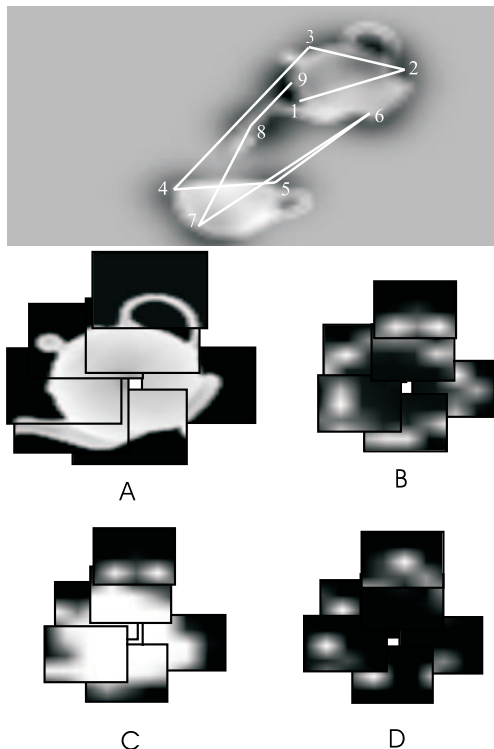


Figure 2: Top:Scan Paths, Bottom:Multiple Fragments of Teapot Object (A) Full visual buffer (B) Edges (C) Region/Boundary (D) Texture

## 5. COMPUTATIONAL THEORY FOR THE FRAGMENTS APPROACH

Our model builds representations from fragments, which in the studies reviewed from the neuroscience literature, correspond to features of objects (e.g. localised parts of objects). For example, in our teapot example, fragments are the features of the teapot such as the spout, handle, bowl-shaped regions at the bottom of the object and so on. These popu-

late a measurement space. An RBF module is tuned for one object, based on its fragment representation.

### 5.1. Basis Functions

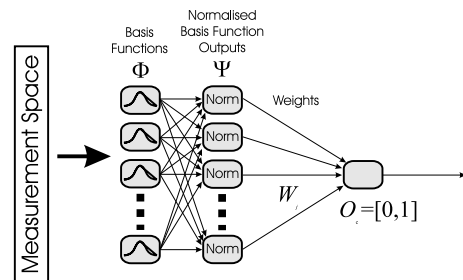


Figure 3: Design of an RBF Module

The first issue is to obtain the relevant RBFs to model clusters in measurement space. Basis functions  $\Phi$  were established using the technique describe in [24]. For brevity, we only outline the technique – full derivations can be found in [15]. An advantage of using this method is that both the widths and the centres of basis functions are learned simultaneously. We use only the unsupervised component of [24] (the GART module). Figure 3 summarises the model implemented for each object’s RBF module.

As the model selects fragments using the attentional system, GART basis functions are established. The best-match basis function is competitively selected by a discriminant function, based on the posterior probability of basis  $j$  being the category for  $x$ . If the fragment also matches the winning basis function’s distribution properties, the basis function is further tuned to include the new fragment. If no winning, suitably matching basis if found, a new one is instantiated. Basis function training proceeds until a quantisation error measure asymptotes.

### 5.2. Complete Modules for Objects

Rather than a linear sum of weighted basis function outputs [1], we preserved the probabilistic formulation introduced in the RBF training. By normalising the discriminant functions  $\Phi$ , we can recover posterior probabilities that any RBF is coding for the presented fragment – see [15] for full details – resulting in an activation vector of probabilities  $\Psi$ . These feed to a weighted sum node, which uses a sigmoid transfer function. Like Edelman’s system, we desire the module to output close to unity for an example of the object, zero otherwise. This is achieved by gradient descent using a cross-entropy error function. Training of these weights is a simple task, with learning rate set inversely to the proportional to the number of training presentations [25]. The result is that the output  $O_c$  for an RBF module is propor-

tional to the probability that the fragment presented is an example of object  $c$ .

## 6. A CHORUS SYSTEM

Recall that the representation space is defined by the number of object-tuned RBF modules present in the system, and can be seen as a vector space  $\mathcal{R}^N$  where  $N$  is the number of object-tuned modules. The representation of a fragment  $\mathbf{x}$  is therefore  $\mathbf{r}_\mathbf{x} = \langle O_1, O_2, \dots, O_N \rangle$  where  $O_c$  is the scalar output of object-tuned RBF module  $c$ .

We trained a complete system on 9 static objects and also, one module was trained on examples of liquid flows where motion was present. In total, some 80 fragments per object were generated by varying the parameters of the attentional system: the decay of inhibition  $\lambda$  was set at 0.25, 0.175 and 0.0875 and the width of the inhibition  $\sigma_w$  was varied between 0.25, 0.5 for each value of  $\lambda$ .

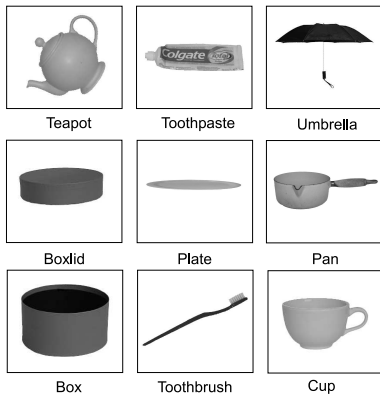


Figure 4: Example Objects

To test the hypothesis that the combination of Chorus-of-Fragments and the attentional system coupling could support object representation, we first tested the system to see if a distributed representation had been formed for the objects. This is signified by relational properties being coded for by activations of different RBF modules in response to any fixated fragment representation. Objects used in the videos are shown in Figure 4. In Figure 5, we tested two visually similar objects (that is, fragments of either would likely coincide in measurement space because of similar rounded edges and texture properties). It can be seen that a distributed representation, usually dominated by the correct module, has been learned.

A similar test was performed with unseen data: a video showing liquid pouring from a teapot into a teacup used in training, but where the attentional system used different parameters (inhibition width and decay) which results in different fragments being extracted through fixation. This represents a real “online” test of the model’s performance and

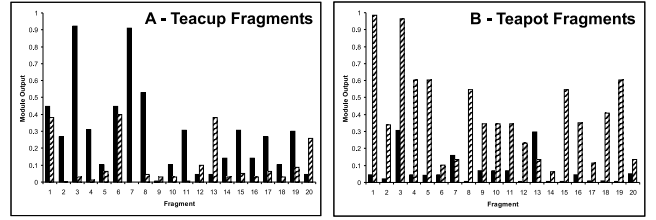


Figure 5: Example outputs for Teacup and Teapot Modules. Black bars are the outputs for the teacup-tuned module and hashed bars represent the teapot-tuned module. (A) The responses for 20 fragments of a Teacup; (B) The responses for 20 fragments of a Teapot

the coupling of learned representation with the dynamics of the attentional system. Figure 6 shows the results, sorted by object.

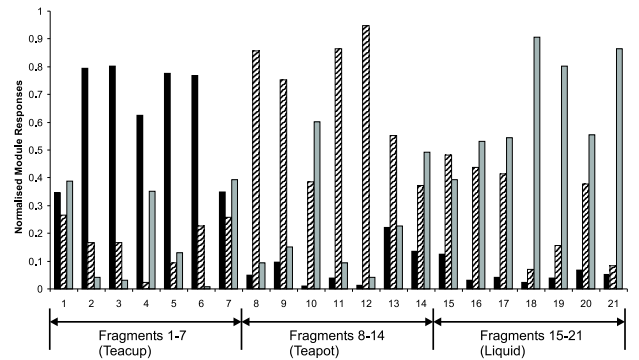


Figure 6: Example outputs for Teacup, Teapot and Liquid Modules for Unseen Video for 7 fixations on each object, sorted by object. Black bars are the outputs for the teacup-tuned module, hashed bars represent the teapot-tuned module, and grey bars the liquid module.

## 7. DISCUSSION

We have shown that by coupling a Chorus system (but which uses fragments rather than whole object views) with a selective attention mechanism provides a means of producing the distributed object representations described by [1, 3]. Because information about fixation and attention is available (our simulated dorsal stream) and object representation (ventral) proceeds by fragments dissociated from retinotopy, we have the basis for an integrated “scene geometry” system. We tentatively propose that the output of a Chorus-of-fragments system corresponds to the properties of the perirhinal cortex, which connects with the entorhinal cortex which receives inputs from both the dorsal and ventral systems [26]. Further work will explore this possibility.

## 8. REFERENCES

- [1] S. Edelman, *Representation and Recognition in Vision*, MIT Press, 1999.
- [2] D.L. Sheinberg, and N.K. Logothetis, "Noticing Familiar Objects in Real World Scenes: The Role of Temporal Cortical Neurons in Natural Vision", *Journal of Neuroscience*, 21(4) pp. 1340-1250, 2001.
- [3] M. Riesenhuber and T. Poggio, "Models of Object Recognition", *Nature Neuroscience (Supplement)*, 3, pp. 1199-1204, Nov., 2002.
- [4] J.I. Nelson, "Visual Scene Perception: Neurophysiology", In M.A. Arbib (Ed), *Handbook of Brain Theory and Neural Nets*, MIT Press, pp. 1024-1028, 1995.
- [5] S. Edelman, "Constraining the Neural Representation of the Visual World", *Trends in Cognitive Science*, 6(3), pp. 125-131, 2002.
- [6] M. Livingstone and D. Hubel, "Segregation of form, colour, movement and depth: Anatomy, physiology and perception", *Science*, 240, pp. 740-749, 1988.
- [7] D.C. Van Essen, C.H. Anderson and D.J. Felleman, "Information Processing in the Primate Visual System: An Integrated Systems Perspective", *Science*, 255, pp. 419-423, 1992.
- [8] J.A. Movshon, E.H. Adelson, M.S. Gizzi and W.T. Newsome, "The Analysis of Moving Visual Patterns" In C. Chagas, R. Gattas and C. Gross, (Eds) *Pattern Recognition Mechanisms*, Springer-Verlag, New York, pp. 117-151, 1985.
- [9] E.C. Hildreth, and C.S. Royden, "Motion Perception", In M.A. Arbib (Ed), *Handbook of Brain Theory and Neural Nets*, MIT Press, pp. 585-588, 1995.
- [10] E.T. Rolls and G. Deco, *Computational Neuroscience of Vision*, Oxford Univ. Press, 2002.
- [11] B.W. Mel, "SEEMORE: Combining Color, Shape and Texture Histogramming in a Neurally Inspired Approach to Visual Object Recognition", *Neural Computation*, 9, pp. 777-804, 1997.
- [12] J. Hegd , and D.C. Van Essen, "Selectivity for Complex Shapes in Primate Visual Area V2", *Journal of Neuroscience*, 20, RC61, pp. 1-6, 2000.
- [13] K.R. Gegenfurtner, D.C. Kiper and J.B. Levitt, "Functional Properties of Neurons in Macaque Area V3", *Journal of Neurophysiology*, 77, pp. 1906-1923, 1997.
- [14] J.V. Stone, "Object recognition using spatiotemporal signatures", *Vision Research*, 38(7), pp. 947-951, 1998.
- [15] D.W. Joyce, L.V. Richards, A. Cangelosi and K.R. Coventry, *Object Representation by Fragments: Toward a Neurocomputational Model of Scene Geometry Analysis*, Technical Report, Centre for Neural and Adaptive Systems, School of Computing, University of Plymouth, 2002.
- [16] B.K.P. Horn and B.G. Schunck, "Determining Optical Flow", *Artificial Intelligence*, 17, pp. 185-203, 1981.
- [17] C. Koch and S. Ullman, "Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry" *Human Neurobiology*, 4, pp. 219-227, 1985.
- [18] J.K. Tsotsos, S.M. Culhane, W.Y.K. Wai, Y. Lai, N. Davis and F. Nuflo, "Modeling visual attention via selective tuning", *Artificial Intelligence*, 78, pp. 507-545, 1995.
- [19] D. Boussaoud, G. di Pellegrino, and S.P. Wise "Frontal lobe mechanisms subserving vision-for-action versus vision-for-perception", *Behavioural Brain Research*, 72, pp. 1-15, 1996.
- [20] J.H.R. Maunsell, "The Brain's Visual World: Representation of Visual Targets in Cerebral Cortex" *Science*, 270, pp. 764-769, Nov., 1995.
- [21] M. Corbetta, and G.L. Shulman "Control of Goal-Directed And Stimulus-Driven Attention in the Brain", *Nature Reviews: Neuroscience*, 3, pp. 201-215, Mar., 2002.
- [22] R.A. Andersen "Multi-modal Integration for the Representation of Space in the Posterior Parietal Cortex", *Philosophical Transactions of the Royal Society of London, B*, 352, pp. 1421-1428, 1997.
- [23] T.R. Vidyasagar "A neuronal model of attentional spotlight: parietal guiding the temporal", *Brain Research Reviews*, 30, pp. 66-76, 1999.
- [24] J.R. Williamson "Gaussian ARTMAP: A Neural Network for Fast Incremental Learning of Noisy Multi-dimensional Maps", *Neural Networks*, 9(5), pp. 881-897, 1996.
- [25] C.M. Bishop *Neural Networks for Pattern Recognition*, Oxford Univ. Press, 1995.
- [26] W.A. Suzuki, E.K. Miller, and R. Desimone, "Object and Place Memory in the Macaque Entorhinal Cortex", *Journal of Neurophysiology*, 78, pp. 1062-1081, 1997.