Available online at www.sciencedirect.com
SCIENCE DIRECT®
Cognitive Systems Research 6 (2005) 333–363
Cognitive Systems RESEARCH
www.elsevier.com/locate/cogsys

A cognitive model in which representations are images

Action editor: John Barnden

Janet Aisbett, Greg Gibbon *

Faculty of Science and Information Technology, The University of Newcastle, University Drive, Callaghan 2308, Australia

Received 22 March 2004; accepted 18 February 2005
Available online 23 May 2005

Abstract

Interpretations of images of the brain are starting to reveal the conceptual tasks in which the person was engaged at the time of imaging. Existing mathematical models can explain the patterns of activity observed in such images in terms of the coherent activity of large populations of neurons, but not in terms of cognition. This paper is an early investigation into how such patterns might provide the internal representations for a cognitive system. Probes, working memories and memories are all represented as images. The accompanying process model describes how attention is set according to the contents of working memory, how attention determines what parts of the probe are memorised, how memories are activated according to similarity to the probe in areas in attention, and how working memory is managed. The model is demonstrated on re-creations of classic simulations of recognition memory and categorisation.

Keywords: Internal representation; Cortical maps; Process models; Attention; Recognition memory; Categorisation

1. Introduction

Neuroimages provide an intriguing but baffling window into human cognition (Poldrack, 2000). Cortical activity captured in such images is not randomly distributed. It is spatially and temporally continuous, at least at the resolution of current imaging. The spatial organisation reflects underlying structure in perceptual input, such as the pitch of a sound, and hints at how higher-level concepts are represented. The temporal organisation hints at how perceptual stimuli are processed and how cognitive tasks are performed. As advances in imaging techniques improve temporal and spatial resolution and reduce noise, there is reason to believe that cortical images will be key to understanding how we think.

Images taken while subjects are engaged in a perceptual or conceptual task have been used to assign a range of cognitive responsibilities to parts

* Corresponding author.
E-mail address: greg.gibbon@newcastle.edu.au (G. Gibbon).

1389-0417/$ - see front matter
doi:10.1016/j.cogsys.2005.02.004

of the brain, in the burgeoning cognitive neuroimaging literature (e.g., Courtney, Petit, Haxby, & Ungerleider, 1998; Duncan & Owen, 2000; Engel & Singer, 2001; Haxby et al., 2001). Attempts have been made to predict from cortical maps what tasks the subjects were engaged in at the time of imaging (e.g., Wang, Hutchinson, & Mitchell, 2003). Models of coherent activity of large populations of neurons have been developed to explain cortical patterns of activity in terms of average neuron firing rates and the like (e.g., Erwin, Obermayer, & Schulten, 1992; Liley, Cadusch1, & Wright, 1999). But there have been few attempts to build these large-scale patterns of activity in the brain into computational models of cognition. Biological computational models – those that attempt to describe how processing is implemented in the brain – almost invariably employ neural nets (Norman, 2003). Such models have been very successful at predicting human performance on memorisation and conceptual tasks (e.g., O'Reilly, Norman, & McClelland, 1998; Kruschke, 1992). However, they are constructed from what are – in biological terms – tiny numbers of rudimentary models of neurons.

This paper develops a formal mathematical model of a cognitive system (Aisbett & Gibbon, 2001) into a computational model of cognition in which information is represented using continuous spatial functions, or images. The system is demonstrated by simulating two classic cognition experiments into recognition memory and categorisation that have previously been explained using attribute vectors as internal representations. The new model is called CIM, for Cortical Image Manipulation.

Many physiological functions can be considered to be points in an infinite dimensional space in which the dimensions are the spatial positions on a cortical layer. The two characteristics of *spatial organisation* and *infinite dimensionality* distinguish representations based on images in general, and cortical maps in particular, from traditional representational forms.

The link between the images in CIM and fields of biological activity is at this stage only by analogy. We do not attempt to map locations in the image plane to locations on cortical layers. Nor do we specify whether the images represent local average neuronal firing rates, or the electromagnetic fields proposed by McFadden (2002) as the source of consciousness, or any other physiological function definable on a surface. CIM is, however, intended to provide the groundwork for future use of such physiological images, or sequences of images. Although CIM is not connectionist, it could be biologically related to connectionist approaches because it is based on coherent activity of large neuron populations.

Before giving an overview of the CIM model, the next section presents terminology and constructs from cognitive modelling in order to motivate our later use of terms. The overview and the formal mathematical definitions follow in Sections 3 and 4.

The question of the type of image intended to be used as input in the modelling is addressed in Section 5: this is important since we do not propose at present to use actual cortical images. In brief, the images are spatially organised, they are used to represent abstract as well as concrete concepts, and they are analogical in the sense of having structure which carries information about the concept represented (Sloman, 1978). For example, any concept associated with magnitude – height for instance – might be represented by a family of radial functions in which the distance from the centre is related to magnitude and the absolute intensity is related to certainty. Representation of magnitude through the standard deviation of a circular Gaussian is illustrated for two values in Fig. 1(a). Fig. 1(b) is a possible representation of a plasticine cube (pictured to the left); the four quadrants going clockwise from the left top, respectively, contain descriptors of size and weight, hue, texture and feel, and shade (Aisbett & Gibbon, 2003). These examples and various properties of infinite dimensional and spatially organised representations are discussed further in Section 5.

Demonstrating how the CIM model performs any cognitive task requires simulation of stimuli and memories. Section 6 describes CIM application to word recognition memory. Shiffrin and Steyvers (1997) simulated words as random vectors, so we simulate them as images composed from randomly located circular Gaussians. Section 7 describes categorisation using the CIM model,

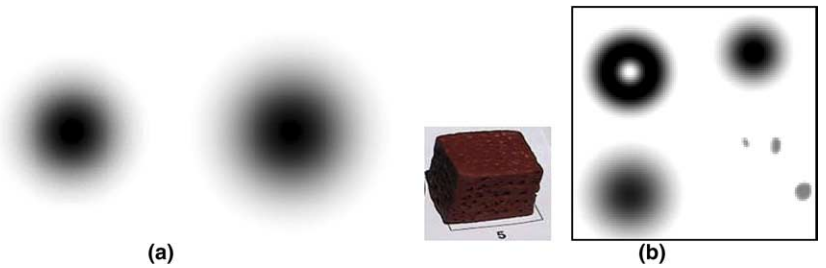

Fig. 1. (a) Radial functions representing magnitude through variance, and uncertainty through intensity. (b) Object and a valid image representation of it. See text for further explanation.

and looks at performance when different families of image are used as representations. Finally, Section 8 points to some of the many directions for future research suggested by this preliminary investigation.

2. Cognitive architectures and representations

The three constructs of *memory*, *working memory* and *attention* are employed in most modelling of the human mind, although terminology varies. Following Baddeley (1986), a Central Executive may be invoked to provide overall control.[1] Other buffers and control mechanisms may be specified to pass sensory-motor material to and from working memory and to react to unexpected stimuli. A mechanism for matching the prevailing situation – as captured in working memory – against stored memories is also usually identifiable, with the closeness of match affecting the activation of memories.

Memory is conventionally divided into declarative and procedural (Ryle, 1949). Tulving and Markowitsch (1998) claim declarative memories have a symbolic representation involving a set of attributes, and can be assigned a truth value. Whatever the case, declarative memory in humans is conventionally subdivided into episodic memories, which allow conscious recall of personal experiences, and semantic memories, which concern factual knowledge. Tulving revised an earlier thesis that episodic and semantic memory exist as separate systems, but a functional division between semantic and episodic memories may exist in the ability to update and extend a semantic memory (Shiffrin, 2003).

Memories are realised as sets of weights in connectionist models. In vector-based representations memories may be noisy or partial copies of the vector representing the original stimulus. In our system they are images that are partial copies of input to working memory.

The notion of *working memory* is controversial (e.g., Cowan, 1988; Miyake & Shah, 1999) with definitions varying as to control mechanisms and the relationship with attention. Working memory makes task-relevant material available over periods in which interfering stimuli may be received. It handles both sensory and semantic information, and enables distinct elements to be composed into units (called object files in visual experimentation (Treisman, 1999)). Some but not all of the contents of working memory become part of permanent memory, and material in it decays or is overwritten unless it is reactivated. We use the term working memory to mean the material directly accessible to cognitive processes, which by the Embedded-Processes Model (Cowan, 1988) consists of activated memories together with material passed to and from sensory-motor buffers.

The notion of working memory has been taken up keenly by computer scientists. Many artificial cognitive systems, including the classic general purpose system Soar (Laird, Newell, & Rosenbloom,

[1] In computer implementations control is often found to be outside the formal model, in the hands of the researcher who organises input data and interprets output data.

1987), have a component called working memory that contains data relevant to the current processing task. In feed-forward connectionist models, working memory is manifest in the activation of input and other nodes.

The blurred relationship between working memory and *attention* was one of eight questions posed by Miyake and Shah (1999) of contributors to their volume of essays on working memory. Baddeley's and Cowan's models of working memory contain a central controller to handle processing and to manipulate attention, which can be envisaged as a spotlight on areas in working memory which defines an *area* or *focus of attention*. Attention is sometimes viewed as a control process, selecting information and operations (e.g., Shiffrin, 2003). From either perspective, there is only *limited capacity* for selecting material, and attended material is only a subset of the material in working memory. Our definition of attention is analogous to the spotlight metaphor, but the direction of the spotlight is a function of the contents of working memory rather than a decision of an Executive.

We adopt the term *probe* from memory research to describe input to working memory from external sources (e.g., sensory buffers), or input generated from the contents of working memory, similar to an object file. A probe is a vector in computational models such as Retrieving Effectively from Memory (REM; Shiffrin & Steyvers, 1997) or the connectionist memory models developed by O'Reilly et al. (1998). For us, a probe is an image.

Many cognitive models are imprecise as to formats used to represent probes and memories. For example, a criticism of dynamical system theory (Van Gelder & Port, 1995) was that the domain of the time-varying functions was hard to pin down (Eliasmith, 1996). Townsend and Thomas (1993) observe that the information processing approach to modelling cognition uses abstract notions of features and attributes which are not explicated. Criticisms of Perceptual Symbol Systems theory (Barsalou, 1999), in which information is represented as activity patterns on modality specific regions of the brain, have been directed at the informality of representation (e.g., responses to Barsalou, 1999). Miyake and Shah (1999, p. 7) found that few of the contributors to their volume of essays on working memory directly tackled the question of how information was represented in working memory and how the representation might vary for different modalities.

Even a computational model trialled with computer simulations – which necessitates precise specification of representations and processes – may be non-committal as to the formats used in the brain. Examples of such abstract computational models are the recognition memory model, REM (Shiffrin & Steyvers, 1997), and the categorisation model, Generalised Context Model (GCM) (Nosofsky, 1986), discussed later.

These particular models and their computer implementations represent information as vectors of attributes, and employ algorithmic processing. Vectors can be viewed as points in a multidimensional space. The notion of *spaces* has been pervasive in psychology, and fits with the view that concepts are defined by sets of attribute values. While the set of dimensions defining such spaces is not spatially organised, each individual dimension usually has a distance defined. Dimensional distances allow the definition of *similarity* between concepts represented as points in the space, which is used in matching memories to probes.

Multidimensional scaling (Shepard, 1966) permits similarity judgements to be used to compute the dimensionality of a conceptual or perceptual space, as well as the location of concepts or percepts in this space. This has been taken as evidence that spaces are valid psychological constructs, not just a computational device. A biological basis for dimensions has been suggested by a lot of research into neural population coding which shows how the output of a large population of neurons can be interpreted as a real value of a dimension (e.g., Abbott & Dayan, 1999).

The biological counterpart of the vector input and output of connectionist models is, in contrast, usually taken to be the activation of a small set of neurons. Representation in neural nets is often described as *distributed*, because concepts are modelled as patterns of activation across a set of nodes. Connectionist models of cognition combine naturally with models of perceptual subsystems

implemented as neural nets, supporting theories that blur the distinction between cognition and perception (e.g., Goldstone, 2003; Tijsseling & Gluck, 2002). In such systems, spatially based representations are used in perceptual processing. Such data are then abstracted into independent high level features, in which spatial relationships are lost, before they are made available for conceptual processing.

While connectionist models dominate much of the current cognitive science and computational neuroscience literature, implemented models are usually hybrid systems, for example utilising rule based control mechanisms. General-purpose cognitive process models such as Soar (Laird et al., 1987) and ACT-R 5 (Anderson et al., 2004) are hybrid or symbolic. Information in symbolic systems is represented as sets of relations on attributes and as rules, which only have spatial reference in specific situations such as the description of geographical or shape-related concepts.

In contrast to the representational formats used in most connectionist and symbolic systems, images are both distributed and spatially organised. This means that a memory can be activated by a partial match with a probe image, supporting inference without the need for rules which relate features, used in symbolic systems, or explicit weighted links, used in connectionist systems. When defined as functions on a patch – that is, a subset of the real plane – images are also infinite dimensional. If features are defined by restricting functions to a connected subset of the patch, then any feature can be subdivided any number of times. Therefore, the feature can be refined as often as required as more is learnt about the environment. Moreover, the subpatches on which features are defined can overlap, allowing interaction between features to be directly modelled. Townsend and colleagues (Townsend & Thomas, 1993; Townsend, Solomon, & Spencer-Smith, 2001) have previously advocated representing information using functions on intervals and subplanes, noting these were infinite dimensional. Linking such function to cortical maps provides a biological rationale for them.

The term "image" has many meanings amongst cognition researchers. Without qualification, it is generally taken to mean a depictive representation which more-or-less preserves relative distances in the object or scene being visualised. In memory research, it may be used synonymously with an episodic or semantic memory. In neurophysiology it is used to denote neural representations of stimuli, as for example, images in olfactory regions of cortex generated in response to odours. It is used to describe spatial organisation in visual memories, which Kosslyn argues have parallels with mental images formed in response to any stimulus (Kosslyn, 2003). The nature of the "images in the brain" reported when we remember things has been a matter of debate over centuries, with Pylyshyn's target article (Pylyshyn, 2002) and the responses to it providing a summary of recent positions.

Despite this overloading of the term, we use "images" to refer to the families of functions defined in Section 4, partly to emphasise that they are analogical, and partly for the connotation with cortical maps. While analogical representation has a long history (see (Sloman, 1978) for discussion of some of it) CIM also has a process model based on pointwise manipulation of the representations to execute a variety of cognitive tasks such as recognition and categorisation.

3. CIM overview

In the CIM model, memories and probes are represented as analogical images, rather than as vectors or arbitrary arrays of numbers and/or truth values. Intensity in the images is constrained to be a member of a family of functions from a connected[2] bounded subset X of $\mathscr{R}^2$ (the plane) to the real numbers, $\mathscr{R}$. The CIM processing cycle is essentially:

$$\begin{aligned}\text{new probe} &\Rightarrow \text{attention setting and memorisation}\\ &\Rightarrow \text{activation of memories onto layers of working memory}\\ &\Rightarrow \text{decay of activity} \Rightarrow \text{new probe}\ldots\end{aligned}$$

The next section describes the formal model more fully. This section presents its main characteristics.

[2] Any two points in a connected subset can be linked by a path that lies entirely in the subset – thus for example the United States is not connected because Alaska cannot be visited from New York without crossing out of the US.

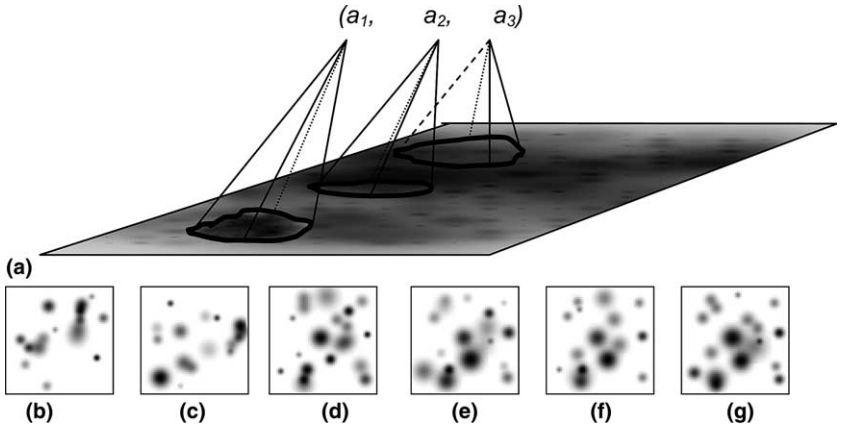

Fig. 2. (a) Values in a conventional vector visualised as average values of patterns on regions of the image plane X. (b)–(d) Examples of simulated memories or probes in which values from a conventional vector have been mapped to standard deviations of circular Gaussians on different locations (means). (d)–(g) Examples of simulated memories taking different values but with fixed centre locations.

The analogues to dimensions in conventional modelling are connected subsets of X, which may, however, be redefined or restructured depending on task or experience. Such subsets may be physiologically realised by what have been called feature maps in sensory-motor areas, for example concerning colour or orientation and so on (Simmons & Barsalou, 2003). Values on dimensions correspond to intensity patterns on the subsets.

While we are not going to specifically associate regions in X with regions in cortex, our model identifies a specific connected subset of X as dealing with indexing.[3] These regions correspond to high level convergence zones proposed by Damasio and used in modelling Perceptual Symbol Systems (Simmons & Barsalou, 2003). Part of this zone or subset of X represents indices to episodic memories, and part represents lexical pointers. These pointers are used to communicate noisily with an external world, through something like the phonological module of Baddeley's Multiple-Component Model (Baddeley & Logie, 1999). Because these pointers are highly identifiable, they may each be associated with a unique region within the index zone.

Fig. 2(a) illustrates how conventional component dimensions from psychological spaces or connectionist models could be derived from images on a plane X. In this example, each component value in the vector (a_1, a_2, a_3) is the integral of image intensity over a region of X specified as a bold outline. A vector dimension is therefore fully specified by specifying a region of X. In this illustration, the regions shown are disjoint, so that the associated vector dimensions are separable. That is, it is possible to independently vary intensity over each of the three regions, and hence to independently vary the components of the vector. This is not the case if two or more of the regions overlap. In vector formulations, dimensions that interact are said to be integral, but there is no way of directly representing such interdependency within the vector format.

To compare any model with results in the cognitive psychology literature, data must be converted to or from the vector format used in reporting experiments (as in Sections 6 and 7). Fig. 1(a) illustrated one way of converting a number to an image, namely, by mapping the value to the standard deviation of a circular Gaussian distribution (or more precisely, to a truncated version

[3] In earlier theoretical work on a related model, this index subset was defined separately from the image space (Aisbett & Gibbon, 2001). There are, however, sensory regions and the like within X which could also be distinguished, so currently we do not make any attempt to formally identify substructures.

of a Gaussian, since the domain must be bounded). An n-dimensional vector can be represented as the composite of n circular Gaussians with centres that might be randomly placed on X or might be fixed, depending on the situation being modelled. When dimensions that are separable in the vector format are modelled this way, their centres must be located so that the overlap of the distributions is negligible. Conversely, interaction between vector dimensions is modelled by siting the centres of the Gaussians so there is overlap, as is seen between many of the dimensions modelled in Figs. 2(b)–(g). Interaction effects are discussed further in Section 5.

Working memory consists of the current probe image, plus images derived from decayed versions of previous probes superimposed with activated memories. A probe image is perceptual input, or combinations of previous probes and activated memories. Capacity is modelled by limiting the number κ of such images that are available for processing at any time. Each image is thought of as occupying a different layer, in analogy to a primary colour in a 3-colour image. Memories activated on the same layer are added together or bound.

Capacity is also limited by limiting the total activation of memories over all layers. This means that activation of a memory works against or inhibits the activation of all other memories.

Biological implementation of the different elements or layers of working memory requires a differentiation mechanism. One such mechanism is the frequency of a carrier wave (of voltage, say) on which activated memories act as an amplitude modulation (Freeman, 1994).

The definition of *attention* is problematic without recourse to an external controller or homunculus, and so in memory models such as REM attention is assumed but not specified (Shiffrin, 2003, p. 345). From the perspective of categorisation, Kruschke (2001) defines attention as a time-varying weight function on input, calculated from the current input and task-related memories including categories. Attention in our model CIM is similarly defined by the contents of working memory, although attention at each layer differs. Attention is used to identify connected subsets of the plane for processing, and these flexibly play the role of dimensions. Biologically, attention might be realised through interference of the time varying signals on X creating short-lived regions of high activity, which nevertheless induce metabolic conditions that facilitate processing not only at that instant, but for a short time after.

The one process model is used in CIM in different cognitive tasks, in particular, in recognition and categorisation. The cycle of activity is driven by the presentation of a new probe, which is either sensory input or the combination of the contents of working memory. The latter is the mechanism through which information is composed or integrated, for instance adding to semantic memories, mentally substituting features on an object, or learning to associate meaning to a word.

Activation of a memory in response to a probe depends on the similarity between probe and memory. Similarity is defined as the integral of pointwise intensity differences, computed over a subset of X that depends on attention and that therefore differs for each layer.

To give a rough feel for the operation of the CIM process model in recognition and categorisation, a version in which working memory is simply the probe can be translated to a 3-layer neural network similar to ALCOVE, Kruschke's, 1992 connectionist implementation of the Generalised Context Model of categorisation. As suggested in Fig. 3, the network has one node per memory in its hidden layer, has input nodes organised into an $n \times n$ image, and has one output node for each category (''recognised'' or ''not recognised'', in the case of recognition). There are $n \times n$ weights – called the *attention image* – moderating the effect of the input layer on the hidden layer. Thus the weight between an input node and a hidden layer node (memory) is the same for each hidden layer node. The weight from a hidden layer node to an output layer node is the similarity between the memory and the image representing the category name, as for ALCOVE (Kruschke, 1992; Eq. (1)). Unlike a conventional neural net, each hidden layer node has its own transfer function which depends on the values of the memory associated with the node. Specifically, the activation of a hidden

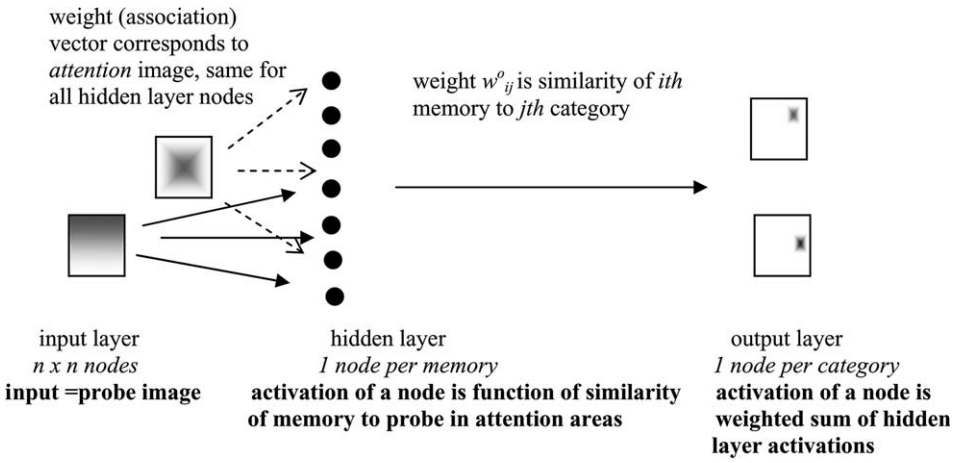

Fig. 3. Translating the process model to an ALCOVE-like 3-layer neural network with one hidden node per memory. All nodes in a layer are connected to all nodes in the next layer. See text.

layer node is a function of the similarity of the attention-weighted probe image to the memory. The activation of an output node is the weighted sum of the activation of the hidden layer nodes. The network learns through adding memories, which alter the output activations that flow from a given probe. Unlike ALCOVE the memory attached to a new node is not a perfect copy of the probe that formed it, but is the subset that was in attention. Unlike ALCOVE, there is no learning of attention, rather attention is re-set to distinctive areas of each probe.

4. Formal definitions

Many stimuli, including sounds, have temporal duration. A full model of a single probe would be as a time varying function over some interval τ. To limit the definitional overhead, the following definitions are, however, restricted to the snap shot view.

4.1. Working memory, memory and attention

Suppose that probes are created or received at discrete intervals T. For simplicity T is taken to be fixed. The probe created or received at time t is a function

$$f_0(t): X \to [0,1] \tag{1}$$

belonging to a family of functions on $X \subset \Re^2$ which are continuous except possibly on the boundaries of some fixed finite partition of X. The value 0 is used to denote "not applicable" or unspecified values, and might be realised as the background noise level in a physiological function.

Consider $f_0(t)$ to be on layer 0 in working memory (the input layer). Suppose each layer $i > 0$ holds an image

$$f_i(t): X \to \Re \tag{2}$$

that was initiated by a probe received at a notional time $t - iT$ and which is now the sum of that decaying memory and memories accumulated in response to it and to any later probes.

Since a new probe is presented every T time units, yet working memory is limited, the contents of an existing layer will usually have to be overwritten by the probe. The easiest approach is to simply overwrite the oldest layer, on the assumption that layers decay into noise over time. Working memory is therefore modelled as a first in–first out (FIFO) buffer of size $\kappa + 1$, for some parameter $\kappa > 0$. That is, after each time interval T, the index to each of the layers 0 to $\kappa - 1$ is incremented and the κth layer is filled with the new probe and

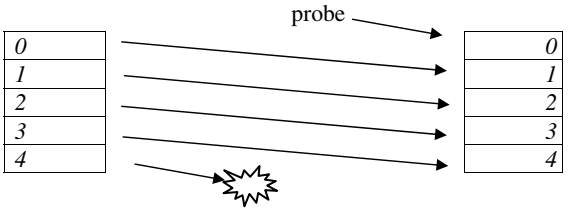

Fig. 4. Working memory as a FIFO buffer. An incoming probe replaces the oldest layer which is relabelled as 0, after the remaining layers are pushed down one slot each.

relabelled as layer 0. Fig. 4 illustrates for a working memory with four layers plus probe.

Next, define *attention* on layer i at time t to be an image

$$A_i(t):\ X \to [0,1]. \tag{3}$$

A major role of attention is to direct what part of a probe image is memorised. Attention over all the layers is assumed to influence the rate of storage of the probe. This makes sense because in a time-limited situation the most relevant parts of the probe need to be stored first, and attention is the only available measure of relevance. If attention at point $x \in X$ is zero on all layers, then the probe image at x might not be memorised no matter how long the interval T. So the rate of storage of the probe image is plausibly determined by the sum of attention $\sum_{i\geqslant 0}A_i(t)$ over each of the layers forming working memory. Specifically, given $x \in X$ and probe f_0 presented at time t, then the value $f_0(t)(x)$ will be memorised during the interval T if and only if

$$\sum_{i\geqslant 0} A_i(t)(x) > w/T \quad \text{for some constant } w. \tag{4}$$

A random variable could be added to the right-hand side of this equation to allow for noisy memorisation. As well, top-down control could be modelled by adding a function $C(x)$ to the left-hand side of Eq. (4).[4]

The portion of the probe that is stored during the interval T constitutes the memorised exemplar or episodic memory.[5] The memory of $f_0(t)$ is therefore the image with intensity

$$f_0(t)(x) \text{ at points } x \text{ satisfying } \sum_{i\geqslant 0} A_i(t)(x) > w/T \tag{5}$$

and with intensity 0 everywhere else. Fig. 5 illustrates storing portions of the probe image in this way in a 5-layer working memory. Fig. 5(a) shows the current probe, as well as the memory of it that will be stored given attention $A_i(t)$ on each of the layers 0–5 shown in 5(b). Fig. 5(b) also shows the contents of working memory at this time, with the probe forming the top layer. The decay of working memory contents over time is indicated by the reduction in overall brightness in each successive layer.

For cognitive tasks that involve identification and discrimination, attention on the probe needs to be triggered by unusual aspects of the new stimulus. Without going to long term memory, the only way to gauge "unusual" is, of course, against contents of working memory. A conceivable mechanism to detect "unusual" happenings is to direct attention to regions of the image plane X in which the firing rate (or voltage or whatever is the modelled biological function) of the probe dominates or is dominated by that of the activated memories. This could be achieved by defining time-varying attention on the probe as the image with

$$A_0(t)(x) = N(t)^{-1}\left| f_0(t)(x) - \kappa^{-1}\sum_{i>0} f_i(t)(x)\right|^r, \tag{6}$$

where $N(t) = \max_{x\in X}|f_0(t)(x) - \kappa^{-1}\sum_{i>0}f_i(t)(x)|^r$ normalises the image, and $r > 0$. Higher values of r lead to a faster drop off in attention. In the following, the exponent $r = 1$ is used.

Attention on each layer $i > 0$ is assumed to be independent of the current probe, since these layers were established in working memory before the probe arrived, and attention is known to have

[4] Alternative capacity limited definitions are possible but are not investigated here. For instance, capacity could be limited on each layer; and if $X(c) = \{x:\ \sum_{i\geqslant 0}A_i(t)(x) > c\}$ for $c > 0$, attention could be defined to be non-zero only on $X(c^*)$ with c^* chosen so that the area of $X(c^*)$ satisfies a capacity-defined upper bound.

[5] As well as episodic memories, memories may be stored combinations of (parts of) other memories, in which case they can be thought of as prototypes or as semantic memories.

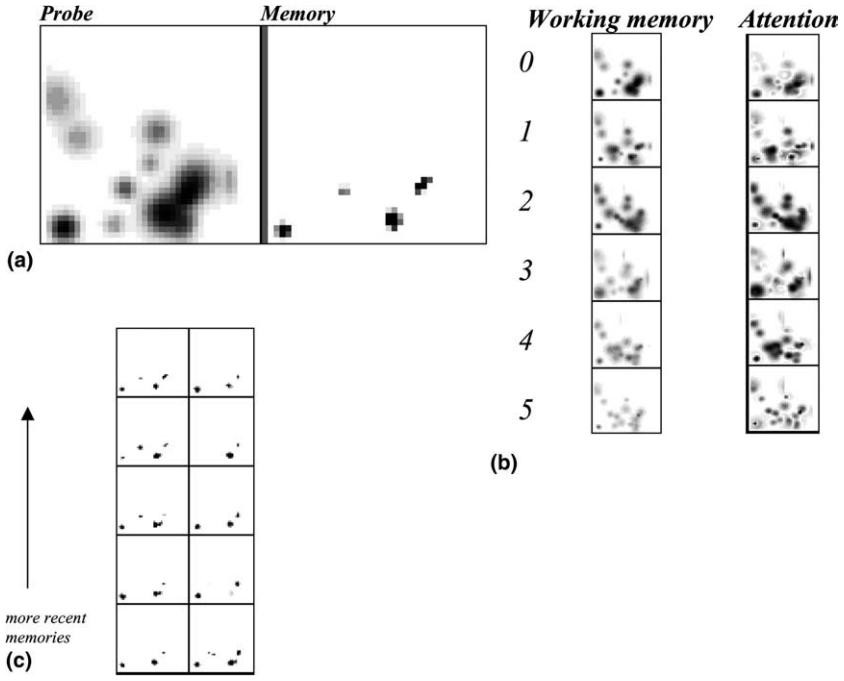

Fig. 5. (a) A probe $f_o(t)$ (left) and memory of this probe (right). The probe is generated from truncated Gaussians, which may overlap as described in Section 5. The memory has four connected areas that act as dimensions, which are partial copies of the original truncated Gaussians or sums of overlapping Gaussians. (b) Left: layers in working memory showing decay of activity. Current probe $f_0(t)$ is in layer 0 (top box) in working memory. Right: Attention layers $A_i(t)$ at the same time. (c) Memories of similar things, as would be encountered in the training phase of a recognition or categorisation experiment. The memory of the probe shown in (a) is at the top left.

some inertia. The layers derive from what were probes at an earlier time, and so attention might be retained from that time, or else be a decayed version. We simply model attention on a layer as the attention image from the time when that layer was the probe:

$$A_i(t) = A_0(t - iT) = N(t - T)^{-1}|f_0(t - iT) - \kappa^{-1}\sum_{i>0} f_i(t - iT)|, \quad i \geqslant 0. \tag{7}$$

Given a cut-off $\delta = w/T$, in which w might physically correspond to ambient noise levels in cortex, the *area of attention* $X_i(t)$ on layer i is defined as

$$X_i(t) = \{x : \ A_i(t) > \delta\}. \tag{8}$$

This area is important in our modelling because activation of memories on the layer will depend on the similarity between memory and probe in $X_i(t)$.

Piecewise continuity of the functions in Eqs. (6)–(8) allows $X_i(t)$ to be expressed as the finite union of connected subsets of X. Likewise the subset $\{x : \ \sum_{i\geqslant 0} A_i(t)(x) > \delta\}$, which defines areas on the probe image which are memorised in period T, can be partitioned into a finite number of connected sets $X^k(t)$. These sets may be likened to dimensions in conventional modelling, as suggested by the simulated memories in Fig. 5(c). The memory in the second box in column 2 has two connected regions, while the memory of the current probe and the previous probe (shown below it) have four regions, and so on.

Given a longer study time T, the size of the regions $X^k(t)$ grow as δ decreases. The number of such sets tends to increase, although there may also be merging of previously disconnected sets. Fig. 8(b) simulates memories formed for two values of the interval T (i.e., different study times), as explained further in Section 6.

Note that if there is no activity in working memory other than the current probe, the area of attention is on regions where the probe takes relatively high values, according to Eqs. (6) and (8). This definition of attention is suited to representations of magnitudes of the sort depicted in Fig. 1(a) but not to direct representations, in which probe activation level is proportional to magnitude.

4.2. Activation

In models of memory and categorisation, a probe's activation of a memory is explicitly or implicitly linked to the *similarity* between probe and memory. In our setting, similarity is derived from absolute pointwise differences $|f(x) - f_0(x)|$. As usual, similarity is a relative notion. If X^* is a finite union of connected subsets of X, then the similarity S of two images f and f_0 in the area of attention X^* is defined to be

$$S(f, f_0; X^*) = \exp\left(-c \cdot \int_{X^*} |f(x) - f_0(x)| \, \mathrm{d}x\right). \tag{9}$$

Here, c is a positive constant that is a free parameter in many models and is usually fitted to experimental data. We are not going to precisely model data so for simplicity fix $c = 1$.

The change in activation of a memory f which contributes to the image in working memory layer i is assumed to be proportional to the similarity between the probe and that memory in the attention area $X_i(t)$ on the layer. Formally this means that if the activation of memory f which contributes to layer i at time t is denoted $\iota_i^f(t)$ and f is not active on this layer until presentation of the probe at time $t - T$, then

$$\iota_i^f(t) = \lambda(t) S(f, f_0(t - T); X_i(t - T)). \tag{10}$$

The normalising variable λ is defined in Eq. (14). As noted, normalisation effectively causes activation of one memory to inhibit activation of other memories.

Decay of activation is, as usual, modelled as an exponential function of time, with decay constant b, say. In the absence of any further stimulation of memory f on layer i, at time $t + T$ activation will have fallen from its level at t according to

$$\iota_i^f(t + T) = \iota_i^f(t) \exp(-bT). \tag{11}$$

Earlier, we said that the accumulation of active memories adds to a decaying probe to form an image on a layer. This can now be formally described, and then λ can be computed. Keep in mind that in the interval T between $t - T$ and t, the notation used to describe the image that was on layer i changes from $f_i(t - T)$ to $f_{i+1}(t)$ because the index is incremented (see Fig. 4). Also recall that for each layer:

(a) The probe activates each memory in proportion to the similarity between memory and probe in the region of attention associated with that layer.
(b) Images decay according to Eq. (11).

Then for $0 \leqslant i < \kappa$,

$$\begin{aligned} f_{i+1}(t) &= \exp(-bT) f_i(t - T) + \sum_{f \in M(t-T)} \iota_i^f(t) \cdot f \\ &= \exp(-bT) f_i(t - T) \\ &\quad + \lambda(t) \sum_{f \in M(t-T)} S(f, f_0(t - T); X_i(t - T)) \cdot f \\ &= \exp(-bT) f_i(t - T) \\ &\quad + \lambda(t) \sum_{f \in M(t-T)} \exp\left(-\int_{X_i(t-T)} |f(x) \right. \\ &\quad \left. - f_0(t - T)(x)| \, \mathrm{d}x\right) \cdot f, \end{aligned} \tag{12}$$

where the sum is over the set $M(t - T)$ of memories available at time $t - T$.[6] It is straightforward to inductively derive an expression for $f_i(t)$ and for the activation of memory f on layer i entirely in terms of the probe and memory images.[7]

[6] There may be a threshold δ so that the sum is over $\{f : S(f, f_0(t - T); X_i(t - T)) > \delta\}$. Such a strategy is suggested by Shiffrin (2003) and others, and might be biologically set by the average magnitude of noise in memory.

[7] $f_i(t) = \exp(-ibT) f_0(t - iT) + \sum_{1 \leqslant j \leqslant i} \lambda(t - (j-1)T) \exp(-(j-1)bT) \sum_{f \in M(t-jT)} \exp(-\int_{X_{i-j}(t-jT)} |f(x) - f_0(t - T)(x)| \, \mathrm{d}x) \cdot f$ and $\iota_i^f(t) = \sum_{1 \leqslant j \leqslant i} \lambda(t - (j-1)) \exp(-(j-1)bT) \exp(-\int_{X_{i-j}(t-jT)|f(x) - f_0(t-T)(x)|} \mathrm{d}x)$.

The variable λ can now be calculated, provided b is known and total activation of all memories on all layers, $\sum_i \sum_f \iota_i^f$, is assumed to be fixed over time. Specifically, if the total activation over layers 0 to κ is a constant Λ, and the activation at the oldest layer κ is assumed to subside into noise in the interval $(t-T,t]$, then from Eq. (11) and then Eq. (10)

$$\begin{aligned}\Lambda &= \sum_{0\leqslant i\leqslant\kappa}\sum_{f\in M(t-T)} \iota_i^f(t)\\ &= \exp(-bT)\left(\sum_{0\leqslant i<\kappa}\sum_{f\in M(t-2T)} \iota_i^f(t-T)\right)+\sum_{f\in M(t-T)}\iota_i^f(t)\\ &= \exp(-bT)\left(\Lambda-\sum_f \iota_\kappa^f(t-T)\right)\\ &\quad+\lambda(t)\sum_{0\leqslant i\leqslant\kappa}\sum_{f\in M(t-T)} S(f,f_0(t-T);X_i(t-T)).\end{aligned} \tag{13}$$

Hence

$$\lambda(t)=\left(\Lambda-\exp(-bT)\left(\Lambda-\sum_f \iota_\kappa^f(t-T)\right)\right)\Big/\sum_i\sum_f S(f,f_0(t-T);X_i(t-T)). \tag{14}$$

4.3. Probes and labels

Probes can derive either from presentation of an external stimulus, or, in the absence of such intervention, from existing working memories. The ability to generate probes from the contents of working memory is necessary to compose elements, and might in some circumstances take precedence over the incoming stimuli.

A probe is generated from the current contents of working memory as follows. Suppose that a probe is presented at time $t-T$, and that by a time t^- just an instant before t the activated memories in the respective layers have accumulated into functions $f_i(t):\ X\to\Re$ and there is no new external stimulus. Then the probe for the time t is defined by

$$f_0(t)(x)=N^{-1}\sum_{i\geqslant 0}\{f_i(t^-)(x)\}. \tag{15}$$

Here, the sum includes the previous probe and N is a normalising factor. The other layers are emptied, and according to Eq. (6) attention $A_0(t)$ is the normalised image $|\sum_{i\geqslant 0} f_i(t)|$. Eq. (15) provides a mechanism for composing or binding concepts represented on the layers, which is needed for tasks such as mental simulation. In cortex, some impulse would presumably be required, for example to physically trigger jumps to a common value in the carrier frequency.

As indicated in Section 3, a subset of X is used for indexing episodic memories and lexical data. This implies that each episodic memory contains a subimage ε which acts as an index. This would record the situation or context in which the memory was laid down as described by Shiffrin & Steyvers (1997). Context subimages recorded at different times would vary, but would tend to be highly correlated when they represent memories of similar times. The image of an incoming phonological or orthographic stimulus β is a lexical pointer defined on a subset $X(\beta)$ of X which is in part uniquely identified with the word.

A probe β activates memories as in Eq. (10) which match or nearly match the label on $X(\beta)$. These are memories that have been associated with the word. Denote the image generated in this fashion by $f(\beta)$. If the word is well known, the response to β is strong. If the word is being learned, there will be only a few memories close to β and the response is weak. In supervised or teacher-led learning (where the word may be an artificial category name) β has to be attached to an image that is being portrayed as its description. In this situation, our modelling assumes that an image representing the name and the image representing the description are activated on different layers then, as in Eq. (15), combined additively on the same layer (at the same carrier frequency, in the biological metaphor). This binds the syntactical label to an image describing the semantics. This process may occur in a phonological component outside the main processing stream, or the composite semantic memory $f(\beta)$ may be already stored. In either case, $f(\beta)$ effectively constitutes the probe generated by the stimulus β.

5. Families of images

5.1. Introduction

Physical time-varying spatial functions in cortex motivated the use of images to represent concepts. It is possible to interpret some fMRI images in terms of the concepts being thought about at the time (e.g., Haxby et al., 2001). However, neither the precision of such imagery nor its association with concepts are at the stage that cortical images can be used to represent objects or concepts. In lieu, representational images have to be generated in the same way that conventional vector representations are generated through simulation or experimentation.

This section concerns this construction of families of images for use in our modelling. It discusses how image representations are capable of naturally encoding information about distributions of feature values, as well as the values themselves. In contrast, in traditional vector representations in cognition, uncertainty has to be modelled in ancillary data structures, such as the circular Gaussians Ashby & Townsend (1986) used to represent perceptual noise, or left as a loose end in the model, as in (Shiffrin & Steyvers, 1997). The image representation provides a plausible explanation of how probabilistic information might be cognitively available.

Unless researchers are directly concerned with perception, models of memory and categorisation utilise categorical or interval alphanumeric descriptions of properties of objects and concepts.[8] Verbal descriptions may be obtained from experimental participants by asking directly for descriptions (from observation or from common knowledge), or asking for a choice to be made from a list of descriptors provided by the experimenter, or asking for similarity comparisons and then applying multidimensional scaling. Or they may be obtained directly from the experimenter's observations or general knowledge.

In direct analogy to verbal descriptions of stimuli, images describing concept instances can be experimentally elicited. Results of initial experiments concerning both objects and abstract concepts are presented in (Aisbett & Gibbon, 2003). In these experiments, the type of images which participants were allowed to construct was constrained by both the aspect of the description (e.g., as "size and weight", or "feel and texture") and by the range of images that could be used to represent each aspect. In constructing image representations, participants were instructed to try to match the perceived similarity between images with the perceived similarity of the concept instances that the images represented. Representations such as that depicted in Fig. 1(b) were the result. This experimental work is still at an early stage. Results in the present paper use images that have been generated by transforming numerical descriptions of concept instances.

5.2. Accumulator coding and its generalisation

There is considerable evidence that both animals and humans have a primitive numerical capability in which magnitude is represented through accumulation of activation, and noise on memories is proportional to the magnitude (e.g., Whalen, Gallistel, & Gelman, 1999). Such an accumulator model of numbers as magnitudes is already familiar to most people through examples such as the thermometer.

Accumulator or thermometer coding of the reals is formally a mapping m from the positive reals to a set of sets that satisfies the condition

$$\text{given any pair of numbers } a, b \text{ with } a < b \text{ then } m(a) \subset m(b). \tag{16}$$

The set to which a value of a feature is mapped can be thought of as a binary image defined on X.[9]

[8] To encourage comparative testing of categorisation algorithms, UCI publishes a collection of diverse datasets containing the category (class) and attributes of sets of entities (Blake, Keogh, & Merz, 1998). Attributes are either numerical or categorical.

[9] Accumulator coding is usually phrased in terms of subsets of the real line $\Re$. The CIM model could be phrased in terms of sets within $\Re$ although the biological rationale and some interaction properties not investigated here would be lost.

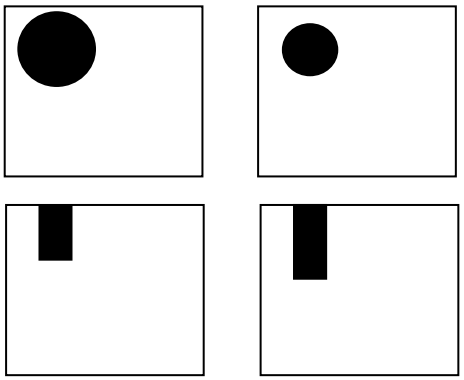

Fig. 6. Examples of accumulator coding of two different values: by radii of discs (top row); by bar length (bottom row).

Examples of accumulator mappings are those that take a magnitude r to an image which is zero everywhere apart from, respectively, (a) a disc of radius r centred at some point in X or (b) a horizontal bar of length r and fixed width with one end fixed in X. Fig. 6 illustrates.

Accumulator coding can be extended to grey-scale images (i.e., images taking values over a continuous finite range of real values) by requiring that whenever a and b are mapped to images I_a and I_b and $a < b$ then

$$I_a(x) \leqslant I_b(x) \text{ for all } x \in X \text{ and}$$
$$I_a(x) \neq I_b(x) \text{ for at least one point } x \in X. \qquad (17)$$

The discrete analogue of this was used in accumulator coding of vector input into neural nets by Kalish & Kruschke (2000). Normalising the range of values over which intensity varies to lie in the interval $[0,1]$ allows intensity to represent probability. This is possible because the numerical values of the feature are represented by a parameterised family of functions on X, that is, by structure in a subimage rather than by intensity magnitude.

5.3. Representing probability density functions

To explore the representation of probability, note firstly that the direct translation of numbers into discs or bar images does not satisfy the requirement that images belong to families which are continuous except possibly on the boundaries of a fixed partition of X. Repeated smoothing of sharp images of bars or discs converge to continuous functions on X. Smoothing can be achieved by accumulation of multiple observations of a stimulus, or of observations gained from multiple examples of a stimulus or concept. In the first case, accumulation models the integration of noisy perceptual responses over an observation period. In the second case, it models the development of a prototype.

To appreciate the role of accumulation in the representation of probability, suppose that feature a takes positive numerical values distributed according to a density function,[10] call it p. As a component in a vector representation, the feature would be represented as the number $\int_0^\infty rp(r)\,\mathrm{d}r$ which is the expected value, or prototypical value. In practice, p is not known exactly, although improved estimates of it are gained with longer observation times (for a constant perceptual stimulus) or with more encounters of exemplars (for building a prototypical value or semantic memory of a higher level concept).

Suppose that the feature a taking positive values in its conventional numerical representation is represented by a family of radial functions centred on c_a in X. Specifically, suppose that if a value of the feature could be known for certain to be r, it would be represented as the image

$$fr(c_a + s\theta) = 1 \quad \text{if } s < r,$$
$$fr(c_a + s\theta) = 0 \quad \text{if } s \geqslant r, \qquad (18)$$

where θ is shorthand for the unit vector at angle θ, and θ ranges from 0 to 2π rad. This is a disc of radius r centred at c_a, and carries no more information than the numerical version. However, when data are accumulated, the image version carries information about the distribution of values as well as the numerical average. In the limit, the accumulated image has intensity function

$$f(c_a + s\theta) = \int_0^\infty fr(c_a + s\theta)p(r)\,\mathrm{d}r = \int_s^\infty p(r)\,\mathrm{d}r = 1 - P(s), \qquad (19)$$

[10] The function p depends on the population under consideration, which as suggested may be multiple noisy observations of the same exemplar, or observations of different exemplars.

where $P(s)$ is the cumulative probability $\int_0^s p(y)\,dy$ and we have substituted from Eq. (18). $1 - P$ is called the complementary cumulative distribution function, or survivor function. When p is a Gaussian the cumulative probability is given in terms of the complementary error function erfc = 1 − erf.

Fig. 7 illustrates how the cumulative distribution function in Eq. (19) is built up through accumulation of observations from encounters with examples of a concept. In this, the representations are smoothed discs. As the number of observations increases, the accumulated subimage tends to the subimage generated by the complementary cumulative distribution function, which for the uniform distribution of this example is a cone.

Mapping numbers to discs does not produce a valid family of images, as noted. An alternative valid mapping would take values (or a positive monotonic function of the numerical values) to the standard deviation of a radially symmetric Gaussian, or an erfc function. The associated variance would reflect the accuracy of the observation, which for perceptual input would never be known with complete certainty. This was how Figs. 1(a) and 5 were generated.

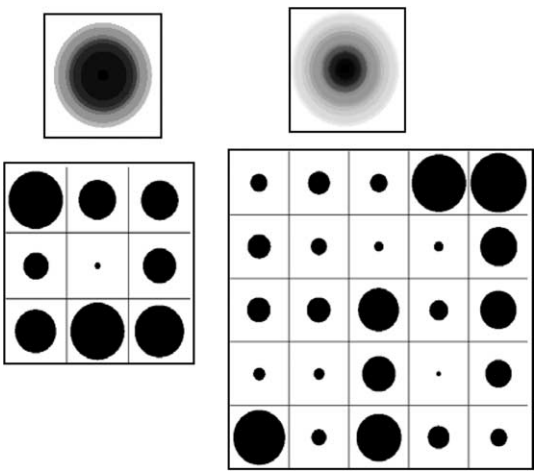

Fig. 7. Accumulation of memories tends to the complementary cumulative distribution function which describes the distribution of the concept. The grid at left (respectively right) depicts nine (respectively 25) memories of instances of a concept. Corresponding accumulated memories are shown above.

When the family of functions representing values of feature a are not radially symmetric, the probability that a has a magnitude greater than y might still be derived from a representation f via a density function $p(y) = N_f \int_0^{2\pi} f(c_a + y\theta)\,d\theta$ for some normalising function N_f.

Not all families of valid representations are able to capture probability distributions through accumulation, of course. In particular, the family that is equivalent to conventional vector modelling cannot capture distributions through accumulation. Consider a fixed partition P_F of X and a family F of functions, each member of which takes a constant value between 0 and 1 on each subset in P_F. The members of F are valid representations of probes, according to Eq. (1). If P_F consists of n subsets, then F is equivalent to a set of n-dimensional vectors in $(0, 1]^n$. If there is exactly one subset in P_F for a perceptual feature a, the representation has no more modelling power than a conventional vector representation – with accumulation, the value on this subset converges to the mean value of a. If the feature a is represented by m subsets in P_F (equivalently, by an m-tuple), uncertainty in an observation of a can be represented by randomly drawing the m values from the distribution that describes the uncertainty. However, the accumulation of a set of observations would not capture the distribution of the observed values, since the value taken on each of the m subsets would converge to the same constant, namely the numerical mean of the observations.

5.4. Using overlap to represent perceptual dimensions which are not perceptually separable

Separable dimensions provide the same perceptual effect when stimuli are varied along one dimension, regardless of values of the other dimensions. The locations in the image plane X used to represent such dimensions need to be selected so that there is no interaction of values, that is, they effectively cannot overlap. In contrast, representations of integral dimensions must overlap to allow changes in values on one perceptual dimension to affect values on the others. At present we have nothing other than separability to guide selection of the locations used to represent perceptual

dimensions. In the simulations that follow, location is therefore chosen randomly if dimensions are not known to be separable, or else they are selected so as to not overlap. If actual cortical maps were being used then the location of dimensions would be constrained according to modality and then by the type of feature.

Interaction effects provide another example of how spatial organisation enhances modelling capacity. If features A and B are represented by radial functions f^A and f^B which overlap for large values but not for small values, then large values of A will raise the perceived probability of a large value of the other feature: formally, $p^A(s) = N_A \int_\Omega (f^A + f^B)(x))\, dx$ where $\Omega = \{x: |x - c_a| = s\}$, N_A is a normalizing constant, and p^A is the density function associated with f^A as in Eq. (19). For example, suppose A is the hue red and B is the hue yellow, and radial magnitude represents brightness in a representation of colour.[11] Then a bright red would appear yellowish even when there was no yellow stimulus, and a bright yellow would appear orange tinged, compared with the pure hues at low brightness levels.

High dimensional representations of the features alone does not provide this modelling capacity. For example, if the feature values are defined by the average intensity on high dimensional vector spaces X_A and X_B with non-null intersection $X_A \cap X_B$, then the features interact. However, the interaction occurs over the entire range of feature values, and is proportional to the number of components in the intersection, as well as to the values of the features.

Representation using images instead of vectors is also enriched by the possibility that noise affects locations of features as well as values. Small shifts in the central location might accompany perceptual noise. Large shifts might be due to changes in perceptual attention altering the probe image presented to our system in response to a stimulus. (Such perceptual processing is assumed to occur before the cognitive modelling stage considered here.)

[11] This paper is not the place to go into the complex representation of colour. Modelling colour remains an important unresolved (and possible unresolvable) issue (e.g., Saunders & van Brakel, 1997).

6. List memorisation, recognition and recall

6.1. Overview

This section reports on a simulation of a list-learning experiment. Our aim is to demonstrate how memories are formed from representations of stimuli using the pointwise operations presented in Sections 4.1 and 4.2, and how the label dimensions described in Section 4.3 are used in simulating responses to lexical input. We also identify an experimental result that might be explained by the effect on attention of the contents of working memory when image representation is used but which could not be explained with conventional representations.

The list-learning experiments of Ratcliff, Clark, & Shiffrin (1990) involved word stimuli presented visually which participants memorise. Later, words which may or may not have been on the learned list are presented to the participant who has to identify whether the word was on the list (“old”) or not (“new”). Shiffrin & Steyvers (1997) developed versions of a model they called REM to explain results from such experiments. REM assumed various distributions of context, visual form and semantic feature values, and assumed the same distributions held for the general word population and the list word population unless the list was supposed to contain unusual words. Context features describing the situation at the time of memorisation were assumed to be specific to the list words. For the sake of simplicity – although it could also be justified in terms of Zipf’s law for natural populations – Shiffrin and Steyvers used a geometric distribution, i.e., $P_i(v = j) = (1 - g_i)^{j-1} g_i$ for some $g_i \in [0, 1]$. Memorisation of any value was assumed to have a fixed chance of error independent of the value, with erroneously remembered values drawn from this same distribution.

In the recognition phase of the list-learning task, a test stimulus is compared with each of the memories that have been activated (above some

threshold level) in order to calculate whether the memory had been a word on the list. REM assumes only list word memories are activated, because the experimental context that is strongly encoded with each of these memories acts as filter. The decision criterion Shiffrin and Steyvers used was the log likelihood value of 1, based on the ratio of the conditional probabilities of observing the test probe if it represents a list word to that of observing the test probe given it represents a word not on the list. The conditional probability of observing a test probe given it was a particular word on the list is in turn calculated as the inverse to the similarity of probe and list word, with allowance made for false (mis)matches due to noise. An overview is in (Shiffrin, 2003).

Our modelling of recognition with CIM follows the REM approach, with three main differences. Firstly, we explicitly model attention, whereas Shiffrin and Steyvers did not try to describe how some features came to be in attention. Secondly, we allow interaction of features, by randomly locating feature on the image plane with consequent random overlap. Thirdly, the recognition process in CIM does not necessarily use semantic memories of word, whereas these were required in REM. Semantic word images are not inconsistent with our model, but could also be formed

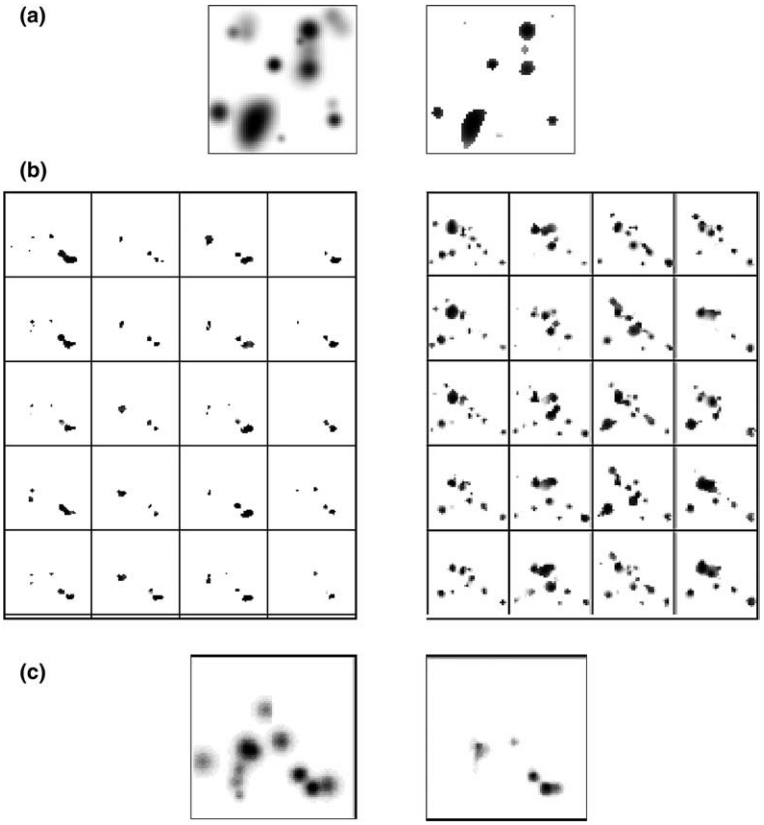

Fig. 8. (a) Example of a probe and its memory, illustrating how the simulation of 20 conventional dimensions with possibly overlapping Gaussians results in a memory of a fewer number of connected subimages, which we call dimensions. (b) Two sets of 20 memories from simulations showing how the same areas tend to be remembered in each simulated word. The set of memories to the right simulates a longer study time (lower attention cut-off) than the set to the left. Note that the two simulations used different locations of the centres of the Gaussians. (c) Recognition phase. Test word (left) and the sum of memories of the 30 words which forms the image f_1 on layer 1 (right). Attention A_1 is the same as f_1.

from the accumulation of episodic memories. We are also non-committal as to whether episodic memories of list words accumulate at the time of memory testing to form a prototypical image of the list, or whether such an image is progressively formed and stored.

In representing a word as an image, values on orthographic or semantic dimensions were modelled as circularly symmetric intensity functions generated by Gaussians. The standard deviations of the Gaussians were the log base 2 of a geometric distribution. Thus the spread of these functions plays the role of the component values in vector modelling. Figs. 2, 5(a) and 8(a) give example of such images.

For each experiment, a set of 20 locations was randomly picked. For each word used in that experiment, a random number of locations (up to maximum 5) were assigned to a "missing list", with random assignment of locations to this list. The remaining locations had Gaussian noise of standard deviation 0.3 pixels added before they were used as centres of circular symmetric functions. The intensity functions were added one at a time to the image representing the word, with the image being renormalised to maximum intensity 1 after each. Repeats of the same word had a "jitter" of standard deviation 0.07 on the centres.

In the simulations reported below, working memory had 5 layers plus probe, and all images were 64 by 64 pixels. Unless otherwise reported, the attention cutoff was $\delta = 0.45$ and the parameter of the geometric distribution was 0.3. The decay constant was 0.2. For simplicity we did not add contextual features to the images, instead we algorithmically simulated effects involving these features as described below. Had context been modelled in the image, it would not interact with the word features and would only vary slightly between words within an experiment.

6.2. Memorisation

Assume that words are presented at spacing T. The process phase is initiated in response to the first stimulus word, presented at time t say. The first orthographic stimulus activates the word pointer image (call it β_1), which with situational context features either forms the probe image $f(\beta_1)$ or retrieves it if it exists as a semantic memory. Since there is nothing in working memory (or at least, nothing task-related and consistent across participants that we can model) attention is directed to neighbourhoods in which $f(\beta_1)$ has highest absolute intensity. The extent of $f(\beta_1)$ stored in the episodic memory of this word presentation depends on the study time and the strength of the maxima according to Eqs. (4) and (6). This is analogous to REM modelling of the number of feature values committed to memory.

The probe $f(\beta_1)$ activates other memories according to their similarity in the subset $X_0(t)$ of X that receives attention in this memorisation process (Eq. (10)). Activation occurs on just one layer, that of $f(\beta_1)$, and so the image f_1 for that layer has, by time $t + T$, accumulated as

$$\begin{aligned} f_1(x) &= \lambda(t+T)\sum_f S(f, f(\beta_1); X_0(t)) f(x) \\ &= \lambda(t+T)\sum_f \exp\left(-\int_{X_0(t)} |f(x) \right. \\ &\quad \left. - f(\beta_1)(x)|\,\mathrm{d}x\right) f(x). \end{aligned} \tag{20}$$

Because of the assumed importance of situational context, $X_0(t)$ largely consists of subsets of the image plane X where the list-learning context is represented in $f(\beta_1)$. Since this situational context has not been included in any previous memory, f_1 will be very similar to $f(\beta_1)$.

Suppose the second stimulus word is presented at time $t + T$ and a word pointer image β_2 is formed. According to Eqs. (6) and (7) attention is directed to neighbourhoods of locally maximal intensity in $f(\beta_1)$ and in the absolute difference image $|f(\beta_2) - f(\beta_1)|$, so these are the areas of $f(\beta_2)$ which are memorised. The interpretation of the first word therefore biases the interpretation of the second word. In particular, if there is enough study time then all the regions of the image plane that were involved in memorisation of the first word will be included in the memory of the second word.

The probe $f(\beta_2)$ activates memories on layer 1 according to their similarity to it in the subsets in attention defined by $f(\beta_1)$. By Eq. (12) this results in an image of accumulated memories:

$$f_2(x) = \exp(-bT) f(\beta_1) + \lambda(t+2T) \sum_f \times \exp\left(- \int_{X_0(t)} |f(x) - f(\beta_2)(x)| \, dx \right) \cdot f(x). \quad (21)$$

This time, $f(\beta_1)$ and $f(\beta_2)$ largely share situational context, so most of the difference between them is on regions in $X_0(t) \equiv X_1(t+T)$ describing semantic and visual form. Setting $y_i = \exp(-bT) + \lambda(t+2T) \exp(-\int_{X_0}(t)|f(\beta_1) - f(\beta_2)(x)| \, dx)$, an argument similar to that above yields the approximation

$$f_2(x) \approx y_1 f(\beta_1)(x) + y_2 f(\beta_2)(x). \quad (22)$$

That is, the image at what we can consider layer 2 is, by time $t+2T$ a linear combination of the images of the two words. The image that was the probe at time $t+T$, viz. $f(\beta_2)$, is by time $t+2T$ both decayed and augmented by activation of memory for $f(\beta_1)$, and so is also a linear combination of the images of the words. This forms the image f_1 at layer 1 at time $t+2T$.

When the third word is presented at $t+2T$, attention on layer 0 goes to regions of high positive or negative intensity in $f(\beta_3) - (f_1 + f_2)$, where $f(\beta_3)$ is generated in response to the new stimulus. As well, attention on layers 1 and 2 is, respectively, on regions in which $|f(\beta_2) - f(\beta_1)|$ and $|f(\beta_1)|$ are relatively large. The context argument means only list words are activated as a consequence of any of the word presentations, an assumption Shiffrin and Steyvers also made for REM. So the activation resulting from the probe $f(\beta_3)$ generates linear combinations of $f(\beta_i)$, $i = 1, 2, 3$ on three layers.

The growth in the number of layers continues until working memory is filled. Then the layer with the oldest (and in our simulations, the weakest) activation gets pushed out to make way for a new layer, as in Fig. 4. At each time interval, and at each layer, attention is directed to a linear combination of the images of the words to date, and activation is in response to similarity to the current probe in such areas. Only list words are strongly activated because of the shared situational context. This process results in word memorisation in which previous list words help determine the parts of the word that are stored, that is, its interpretation.

6.3. Recognition

In the test part of typical recognition experiments, participants are presented a sequence of words, and asked to respond after each word whether it belongs to the learned list or not. For simplicity of exposition, take a 50:50 mix of "old" (list) and "new" (non-list) words.

Shiffrin & Steyvers (1997) assume that the situation at the start of this experimental stage triggers memory of the original experimental context. In our scenario this would occur through an index for the event triggering an image s which is the context description. Attention $X_0(t)$ is then directed at high intensity regions within this context description. This activates memories according to Eq. (12), to form the composite "memory of context":

$$f^c(x) = \lambda(t^* + T) \sum_f \exp\left(- \int_{X_0(t^*)} |f(x) - s(x)| \, dx \right) \cdot f(x), \quad (23)$$

where $t^* + T$ is the time at which the first test word is presented. Because the context is unique to the list-learning situation, and is more-or-less shared by all the list word memories, f^c can be approximated as a sum of the episodic memories of the list words.

When the first test word is presented, attention on this layer is set to areas of high intensity in f^c. Let $X_1(t^* + T)$ denote this subset, which as well as including regions in context areas of X would include some regions representing semantic or orthographic features in which the images from all the list words tended to reinforce each other.

Presentation of the first test word results in an image $f(\beta)$ in a different layer to f^c. As before $f(\beta)$ represents the visual form, the semantic content and the situational context. Attention on the

probe layer is on regions where there are large differences between $f(\beta)$ and f^c.

Fig. 8 illustrates the orthographic and semantic part of the images associated with a list of 30 words. Fig. 8(a) shows a probe and its memory, and Fig. 8(b) shows two 20 word memories from two simulations with different attention cut-offs. Fig. 8(c) shows a test word as the probe and the average of the list words, which occupy layer 1.

Memories accumulate at layer 1 up till the time that an "*old*" "*new*" response has to be made, say at time $t^* + 2T$. Memories accumulate according to their similarity to the image f_0 of the test word in the regions in which the list words tend to have common descriptors, to form the image:

$$f_1(x) = \lambda\left(\sum_f \exp\left(-\int_{X_1} |f(x) - f_0(x)|\,\mathrm{d}x\right)\right)\cdot f(x). \tag{24}$$

The image of the test word, whether β is on the list or not, will not generally be defined in situational context areas in X_1 as context has changed from the list-learning situation to the test environment. So only the regions in which the list words have similar semantic or visual form will contribute strongly to the similarity calculation in Eq. (24).

A response "*old*" is made if the total activation in the region of X that identifies the list-learning situation is large enough. Only list words are activated, which is why there was no need to simulate other memories. The simulations include noise on repeats of a word, so that memories do not exactly match probes even when they refer to the same word. Nevertheless, activation is expected to be greater if β is on the learned list than if it had not been, partly because the sum of exponentials $\sum_f \exp(-\int_{X_1}|f(x) - f_0(x)|\,\mathrm{d}x)$ includes one for the memory of the test word f_0, but also because X_1 has been defined so as to accentuate commonalities in list words. This advantage would be expected to reduce with the length of the list, or when the list has many common words, since then the list words become more like the general word population.

Unlike the list memorisation stage, there is no need to remember a word after a response has been made. So working memory reverts to holding just the situational context, and the scenario above is repeated. (In reality, there would also be interference from previously presented test words that could be easily incorporated into the CIM model because it has working memory.)

6.4. Results

Fig. 9(a) gives rates of true positives ("*old*" response to a list word) and false positives ("*old*" response to a non-list word) from simulations averaged over 100 trials for lists of each of 10, 15, 30 and 60 words.[12] The tests consisted of 20 (resp. 30, 60 and 120) words, half of which were the list words and half of which were generated at random. The behaviour accords with results from human experiments reproduced in (Shiffrin & Steyvers, 1997), and with their simulations, although our simulations show poorer recognition on longer lists. Fig. 9(b) gives analogous results in the case of high frequency versus low frequency words, for a fixed list length of 30 words. Low frequency list words are known to allow better recognition performance than high frequency (common) words, and our modelling reflects this behaviour. Frequency of words was modelled by varying the parameter controlling the geometric distribution from 0.2 through to 0.5. At the high end, performance is just chance.

The most interesting result concerns list strengths. The *list strength effect* refers to the fact that the inclusion of so-called "strong" words which have been memorised more completely – whether through longer study times or through repeated presentations during the memorisation phase – does not hurt recognition of other words on the list which have not had the additional study opportunity (Ratcliff et al., 1990). It is not surprising that "pure" lists in which all words have had longer study times or in which all words are repeated lead to improved recognition performance,

[12] "Words" in the simulations are, of course, randomised images, generated as described earlier.

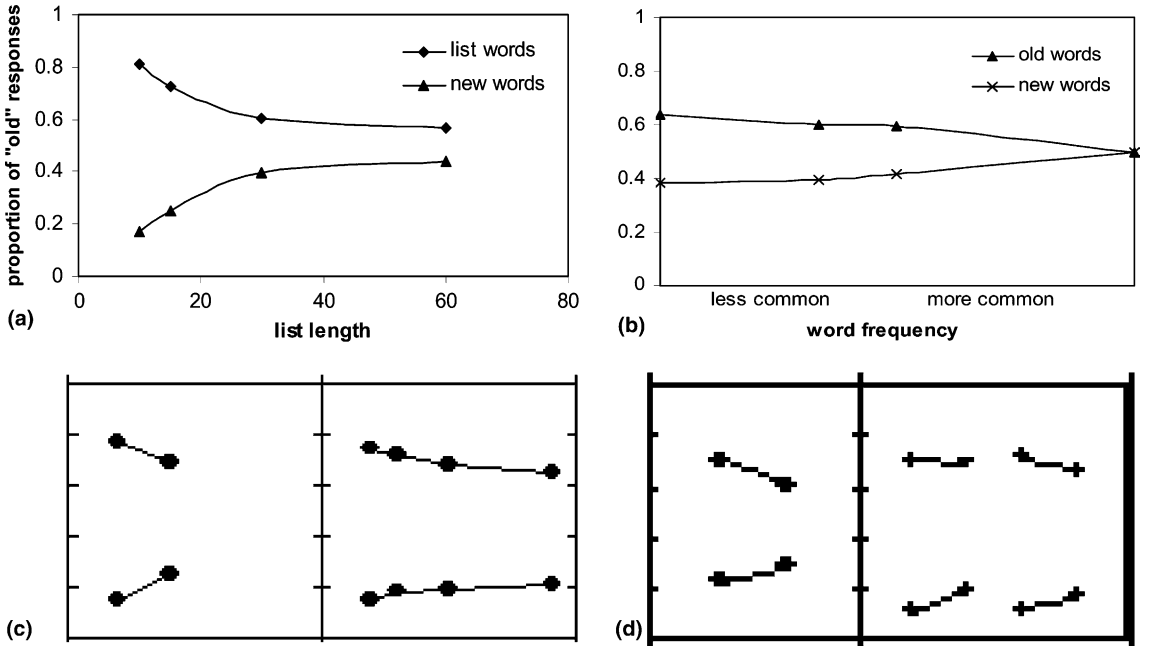

Fig. 9. Word recognition proportions of positive responses to list and to non-list words. (a) and (b) CIM simulation results (a) list length varied (b) geometric distribution varied to simulate different word frequencies. (c) and (d) Experimental and REM simulation results (Steyvers, private communication of data originally presented in Shiffrin and Steyvers (1997)). Experimental results are depicted at left, and REM 5 simulation results are at right. See Shiffrin and Steyvers for a description of REM 5. The vertical axes are proportions between 0 and 1. (c) Experimental list lengths were 10 and 30, while REM simulated list lengths in range 10–80. (d) List length fixed at 30, and the actual environmental base rates of words varied, or the geometric distribution varied.

but the list strength effect on mixed-strength lists is not so intuitive. Ratcliff and colleagues found that the number of false positives decreased pretty much linearly as list composition was varied from all weak words, half and half, to all strong words while keeping the total number of distinct words constant. Likewise, the number of false negatives (failure to recognise a list word) decreased pretty much linearly. Within a mixed list, recognition performance was better on the strong words than on the weak words, but was poorer on strong words from mixed lists than from pure strong lists, and slightly improved on weak words from mixed lists compared with these from pure weak lists.

In explaining such results, Shiffrin and Steyvers argued that the strong words have more features stored and so are more differentiated from the weak words on mixed-strength lists, improving overall performance. Shiffrin has used this result to argue the need for an accumulated semantic memory of the word being formed during a trial, as well as the exemplars (Shiffrin, 2003). However, contrary to the experimental data Shiffrin and Steyvers reported (their Fig. 5) that the REM models' recognition proportions for weak words were slightly decreased using mixed lists, as compared with pure weak lists; and were very similar for strong words and for non-list words whether using pure strong or mixed lists.

Longer study times are simulated in CIM by increasing T or equivalently decreasing the cut-off δ in Eqs. (4) and (5) – in the results reported here, from 0.45 to 0.4. A lower cut-off leads to larger regions being memorised, as illustrated in Fig. 8(b). The larger attention area which is gained in the longer study periods is retained in working memory for five subsequent presentations, and thus in mixed-list conditions can to some extent improve memorisation of words which do not have the longer study time. This explains the upper

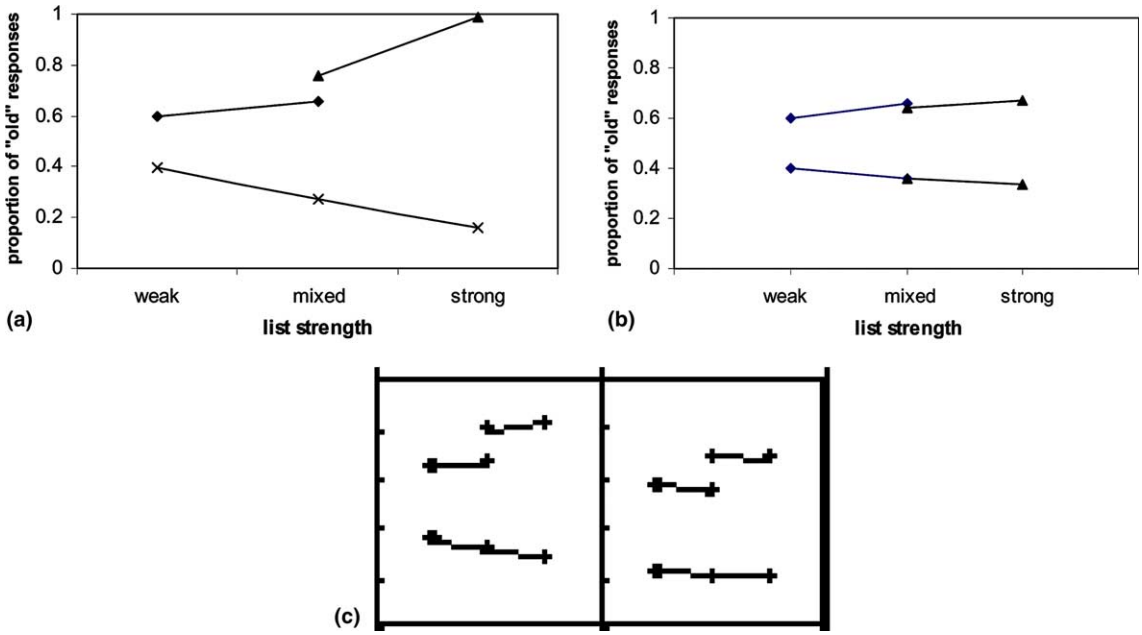

Fig. 10. Effects of list strength on word recognition – upper plots: proportion of positive responses for weak words from pure weak lists, weak words from mixed lists, strong words from mixed lists and strong words form pure strong lists; lower plots: proportion of false positives for pure weak, mixed and pure strong lists. (a) CIM simulations of strengthening through longer study time (b) CIM simulations of strengthening through repeating words. See text. (c) Experimental data (left) and REM 5 simulations (right) for list length 40 (Steyvers, private communication of data originally presented in (Shiffrin and Steyvers, 1997)). Longer study time and repeat words were not reported separately.

plot in the graph in Fig. 10(a), which refers to results simulating lists of 30 words. In the mixed strength condition, performance was better on strong list words than on weak words (76% compared with 66%), but still weaker than when all words had the long study time, when almost all list words were recognised. Performance on strong words was affected because the area that was memorised was reduced in the mixed condition. New words on the list tended to be easier to distinguish on large attention areas as chance match over these areas is less likely. This explains the lower plot in the graph in Fig. 10(a).

Repeated words were simulated as two presentations of the same probe, one after the other, leading to two memories of the same word that might differ with attention. Overall, there is an increase in the area in attention because of the repeats. This led to an increase in recognition of list words even when they were not repeated – in fact in these trials, recognition was slightly better for non-repeated words than for repeated words in the mixed condition. The effect of one repeat of each word is not as great as the change in attention cut-off from 0.45 to 0.4.

In comparing CIM with REM-type modelling, two points should be noted. Firstly, the CIM process model allows earlier states of attention to affect later ones, thus allowing weak words to affect memorisation of strong words and vice versa. This provides a simple and elegant mechanism for achieving the list strength effect which was not available in REM. Secondly, in conventional numerical modelling, attention determines whether features (values on dimensions) are memorised or are ignored in an all-or-nothing way. Therefore, an increase in study time can only affect performance if there are more features or dimensions to be studied. In our analogical modelling, however, increasing attention allows for more

detail about values to be memorised, as well as allowing possible new features to be considered. Therefore, increasing the area in attention can result in a list strength effect even if other factors had restricted study to a single dimension.

7. Categorisation

7.1. Overview of modelling

This section reports on simulations of a categorisation experiment using different families of functions to represent the stimuli. Our aims are to show CIM can produce results comparable to conventional numerical modelling on another important cognitive task, and to explore how using different families of representations affects performance.

In teacher-led (supervised) categorisation, incoming stimuli are provided in a training environment in which a category name is associated with each stimulus, and hence with its internal representation. After the training phase, there is a test phase in which stimuli are presented and subjects nominate a category without receiving feedback. Experiments often vary stimuli along a few physical dimensions such as brightness and size, and attach artificial category labels to disconnected regions in the stimulus space. The way that subjects categorise stimuli not encountered in training is potentially revealing of perceptual similarity between stimuli, and of how different dimensions are manipulated in the decision process, for example, with differential weightings.

In a typical categorisation experiment involving artificial categories, initially no semantic information is known, so nothing is attached to the category name. The training phase builds up knowledge of the category. Only the parts of the probe in attention are memorised, as usual, but the correct category name is attached to this memory – that is, the relevant name image is bound to the memory when feedback as to the true category was provided. The process of giving a response – nominating "*category A*" say – is assumed to have no effect on working memory. (If there is no feedback as to correct category, there is therefore no opportunity to bind a category name to a stimulus. The assumption would be modified in detailed modelling.)

When the test phase commences, memories are activated in response to the category names. These memories might be bound to the names in a phonological component at this time; or a semantic memory of the category might already exist. In either case, we assume that in the test phase each category name image with its accumulated memories forms a layer in working memory.[13] Attention is directed at these layers and at the probe representing the current test stimulus according to Eqs. (6) and (7). If there were more categories than layers in working memory then memory management would be more complex.

The probability of a response – naming of the category of a stimulus – is determined by the total activation of memories in the region corresponding to the category names, compared with the total activation over all categories.

7.2. The experiments

In a classic paper, Nosofsky (1986) reported experiments with two participants who engaged in a very extensive set of identification and categorisation trials involving visual stimuli. The identification phase asked participants to assess similarity of stimuli then used multidimensional scaling to establish the location of the stimuli in the participant's psychological space. The categorisation experiments involved many repeat trials that enabled precise estimates to be made of the probability of a participant categorising a stimulus a particular way (and the most probable decision

[13] Order effects in testing stimuli have been observed in some experiments, and could presumably be captured by our model which has working memory. More detailed modelling in the test phase might also allow working memory to contain images of stimuli found to be surprising, in the sense that they were categorised confidently but the training feedback did not agree with the response. This potentially useful extension would again be relatively straightforward to implement in our model, because of the explicit provision for working memory.

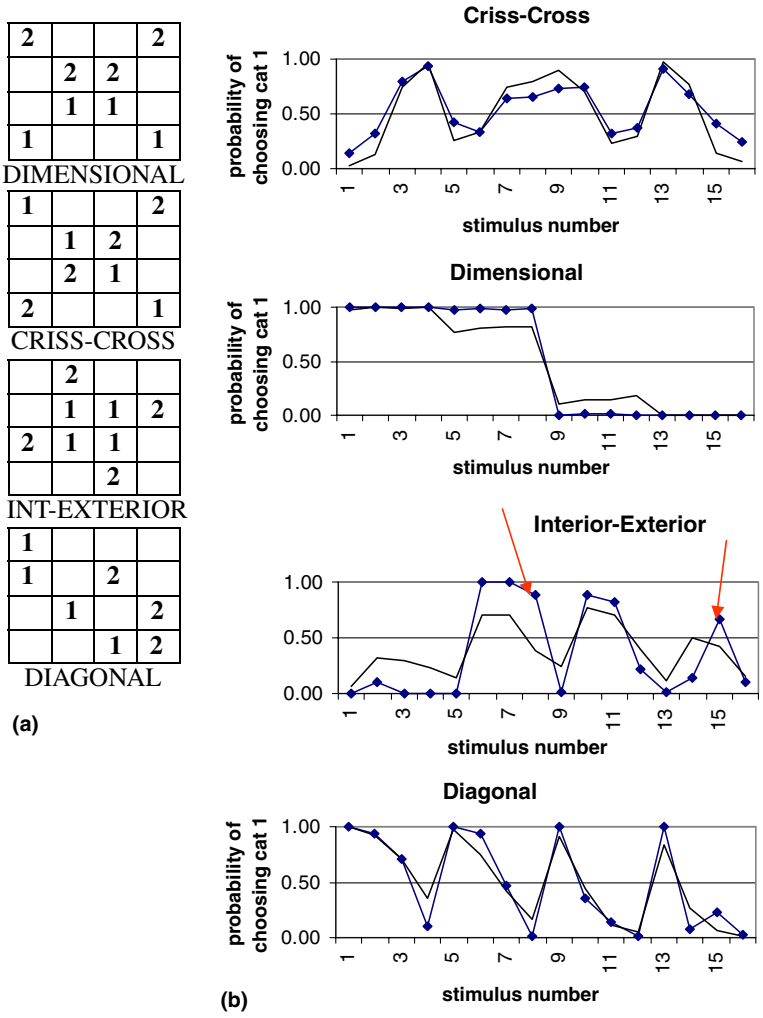

Fig. 11. (a) Structure of the four two-dimensional stimuli sets. For interpretation, see text. (b) The chance of choosing category label 1 given a stimulus, where the stimulus labelling is as in Nosofsky (1986). Actual experimental results reported in (Nosofsky, 1986) for Participant 1 are shown as lines without data points. Averages of 10 simulations using Gaussian modelling with fixed parameter settings across all structures are shown as lines with data points. The two stimuli for which simulated and experimented category labels differ are pointed out as arrows (both occur in the Interior-Exterior structure).

for one participant was not necessarily the same as for the other participant). Distances in psychological space were plugged into the GCM of categorisation.[14]

All of Nosofsky's trials involved two categories, with stimuli varying in two dimensions. Each stimulus could take one of four possible values on each of the dimensions. Four category structures were investigated, called **Dimensional**, **Criss-Cross**, **Interior-Exterior**, and **Diagonal**. These are summarised in Fig. 11(a) as four grids. The 16 positions within each grid indicate the values that each stimulus takes on the two dimensions, and the number at

[14] Nosofsky presented an extended version of GCM that took into account the effects of test stimuli. For simplicity we do not consider the effects of memories acquired during testing.

that position indicates the category named in training for that stimulus. The absence of a number means the stimulus is encountered in the testing phase only. Its assignment to a category is therefore subjective. The stimulus numbers on the horizontal axes of the plots are as in (Nosofsky, 1986) and correspond to increasing values in dimension *1* while keeping dimension *2* constant, then increasing values on dimension *2* (that is, going from the bottom left to the top right of the grid tables). The stimuli that had been encountered in training can be read off from the grid boxes in Fig. 11(a). For example, for **Diagonal** these are the stimuli numbered *3*, *4*, *6*, *8*, *9*, *11*, *13* and *14*. Variability in the perceptual values estimated by Nosofsky, as well as noise in our modelling, means that the differences between category averages are not zero even for a structure such as **Interior-Exterior**.

The GCM gave an extremely close match to the experimental data after fitting just three free parameters for each categorisation experiment: a distance scaling on the exponential in the similarity function (cf. Eq. (9)); a relative weight on the influence of the two dimensions; and a relative modulating weight on each category's activation. The CIM simulation does not try to improve on this match, rather our aim is to illustrate how the spatial representation and the process model developed in Section 4 can be applied to the key task of categorisation.

7.3. CIM simulations

In the simulations reported below, a stimulus was represented by a notional subimage representing the label, and the actual simulated image representing the perceived dimensional values for Nosofsky's Participant 1. Dimensions were assigned disjoint subsets, simulating independent dimensions. Two sets of simulations were performed using representations from Gaussian and erfc families, respectively. A third set was based on representations from a family of functions in which a perceived dimensional value y was represented by a patch of 400 pixels with intensities drawn at random from a uniform distribution with mean y. Examples of representations of the smallest and largest dimensional pairs are shown in Fig. 12(a) for random patch modelling as well as for Gaussian modelling.

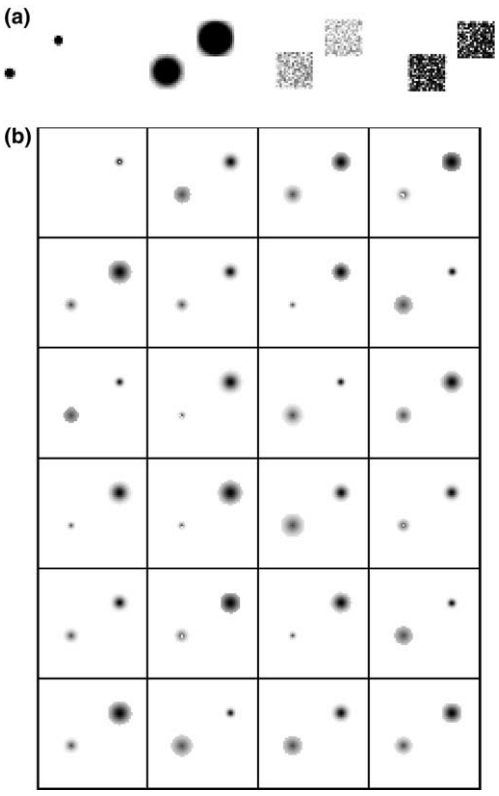

Fig. 12. (a) Simulated probes for smallest and largest pairs of dimensional values. Left, Gaussian modelling; right, random patch modelling. (b) 24 memories of stimuli from training phase, erfc modelling.

All training phase probes were assumed to enter a working memory of four layers + probe, with a decay constant of 0.7. The cut-off on attention was set low (0.2 for Gaussian simulations, 0.3 for erfc simulations, 0 for random patch[15]), corresponding to a long study time. Otherwise,

[15] CIM modelling selects points at which the total attention is above the cut-off (Eq. (4)). This makes sense when input intensity relates to probability, but not when it relates to value. Not surprisingly, much better results are obtained with random patch representations when attention cut-off is set to 0 (all pixels in each patch are used), and the parameter c in the similarity function (Eq. (9)) is varied. These results are presented here.

parameters were set as in the memory simulations reported in Section 6.

32 training examples were presented in each simulation run, selected randomly without replacement from a set of four copies of each of the eight training stimuli. Examples of memories generated from this phase using erfc modelling of stimuli are shown in Fig. 12(b).

Each of the 16 stimuli was presented in turn, and the total activation of each of the category names over each of the active layers of working memory was computed. From this, the probability that category 1 would be selected in response to a given stimulus was derived.

The average of this probability over 10 simulations using Gaussian modelling is reported in Fig. 11(b) against the experimentally observed probabilities reported by Nosofsky for Participant 1. Our results were obtained by varying only one parameter – the attention cut-off – which was then fixed between experimental conditions. The experimental data were fitted better by Nosofsky using conventional modelling and varying three parameters. Nevertheless, the simulations are plausible explanations of the experimental results given there was no attempt to tune parameters. The two cases out of the 52 for which the simulated probability is less than 0.5 and the experimental probability is greater than 0.5 (or vice versa) are pointed out as arrows in the **Interior-Exterior** structure plot.

Erfc modelling in stimulus representation gave similar results, with again two cases of incorrect categorisation, both occurring with the **Interior-Exterior** structure (see Fig. 13). In contrast, the best trial fit obtained on varying the similarity weight for random patch modelling gave five incorrect classifications. Moreover, the weighting c that we found gave the best fit for this structure did not fit as well on other structures, especially the **Diagonal** shown in Fig. 13.

In conventional one parameter exemplar-based modelling, an observation with pair of attributes

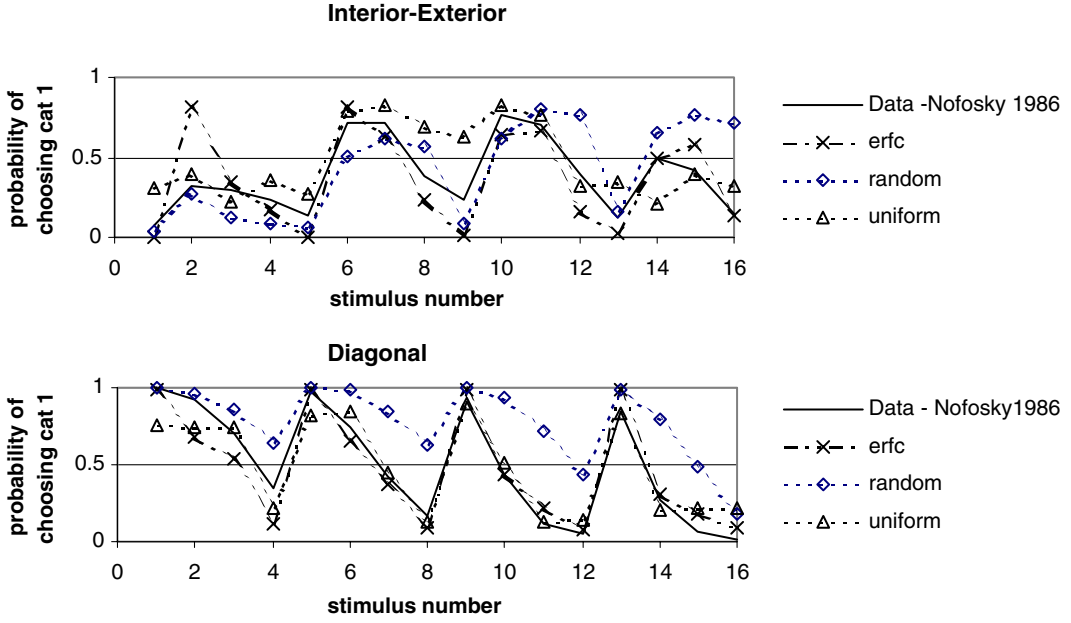

Fig. 13. For **Interior-Exterior** and **Diagonal** structures, actual results for Participant 1 (Nosofsky, 1986) compared with average of 10 simulations using erfc and random patch modelling. The average of 10 simulations using exemplar based numerical modelling with one free parameter is the plot labelled "uniform". Parameters in simulations were fixed at approximate best fit for **Interior-Exterior** structure except for the numerical modelling. For **Interior-Exterior**, experimental and simulated categorisations differ on two stimuli with erfc modelling (stimuli 2 and 15) and numerical modelling (stimuli 8 and 9) and on five stimuli with random modelling (stimuli 8, 12, 14, 15 and 16).

(x, y) is classified as category 1 if and only if $\sum_{j\in C1} \exp(-c(|x - x_j| + |y - y_j|) \geqslant \sum_{j\in C2} \exp(-c(|x - x_j| + |y - y_j|)$, where *Ci* is the set of exemplars (x_j, y_j) in category *i*. We ran simulations using this model (equivalent in our modelling to using functions that are constant over the entire area of a dimension, and setting attention cut-off to zero). Results are shown in Fig. 13; for these, *c* was varied between simulations to give approximate best fit. There are again two misclassifications for **Interior-Exterior**.

The sum of squared errors over the 10 runs for erfc, random patch and numerical modelling was 0.23, 1.81 and 0.13, respectively, for **Diagonal** and 0.45, 0.82 and 0.55, respectively, for **Interior-Exterior**. So the results of CIM modelling with erfc, that is, with a generalised form of accumulator coding, appear comparable to conventional numerical modelling when the same number of parameters are involved.

8. Discussion

The research reported here is an initial exploration of a model of cognition, CIM, in which concepts are represented as continuous functions on a patch (that is, on a connected subset of the real plane). Our opening argument was that the model was positioned to use concept representations based on high resolution neuroimages collected from subjects when they thought about concepts, and these neuroimages would reflect physiological fields. We accepted that current images of the brain, of any modality, were too noisy and too low resolution and too little understood for this purpose. We nevertheless saw the modelling as being biological, despite its limitations, in the same way that connectionist models are described as biological despite the crudeness of computational neurons (Norman, 2003).

Models of cognition must specify both the representation and the process model, and the central role that *similarity* of representations plays in CIM processing ties the two together. Nevertheless, the contributions of the CIM representational form and its process model need to be considered separately. We admit to being more committed to the former than to the details of the latter. Whatever the final form of the process model, it should identify working memory and have explicit mechanisms for setting attention.

The process model in CIM is based around the presentation of probes into a working memory that could hold remnants of previous probes. Probes derive from sensory sources, or from the contents of working memory. Attention is directed to areas of a probe that are different to the average contents of working memory. The parts of a probe that are memorised depend on attention and on study time available. So only fragments of a probe are memorised, as illustrated in Figs. 5 and 8, and the spatial extent of the fragments depends on time before interruption.

The memorised fragments of a stimulus can be interpreted as the record of its distinctive features. The nature of features was only briefly discussed in this paper. However, the new perspective on features that is opened up by the use of continuous spatial representations is one of the most interesting aspects of the new model. The word list simulations showed how mapping from a vector of 20 real numbers to an image composed additively from 20 subimages defined by the numbers does not necessarily mean that there will be 20 features. Rather, overlap of domains of definition cause interaction effects that may create distinctive regions, viz. features.

The ability to represent interaction of dimensions is an important contribution of the continuous spatial representation. In psychology, dimensional interaction has tended to be discussed in terms of integral dimensions, with the example of colour perception playing a leading role. The nature of the interaction is generally captured only in the order of Minkowski metric which is fitted to experimental data (e.g., Shepard, 1966). The assumption that stimuli lie in Minkowski spaces, or indeed in a finite dimensional space, is criticised by Townsend & Spencer-Smith (2003). Spatial representations provide a visualisation of interaction effects in infinite dimensional spaces. By looking at effects that come from biologically plausible image structures, there is the potential to better understand even such a well-studied domain as colour.

A related aspect is dimensional restructuring and reorganisation. Schyns, Goldstone, & Thibaut (1998) argue that humans do not come equipped with a predefined set of features from which all future concepts are constructed, but rather, existing features are refined, existing features are combined in new ways, and entirely new features are created. Describing such restructuring of a conceptual space is cumbersome in conventional vector formats. The temporal evolution of dimensions has to be carried in an ancillary representation to allow memories recorded under one set of dimensions to be compared with memories or probes specified as vectors after a dimensional restructuring. Spatial representation in contrast provides a natural setting for changes because the set X on which all memories and probes is defined is stable.

An obvious disadvantage of spatial representations over numerical representations at this stage is the number of ad hoc decisions on the form of the images needed to simulate any cognitive task. This is related to their lack of parsimony.

CIM was presented in terms of continuous real valued functions rather than the finite approximations that are used in computer implementations of the theory. This is analogous to length, say, being represented on a continuous scale even though any measurement of length will be to a finite resolution. Magnitudes and probabilities of magnitudes are key concepts in representation that are naturally set in continuums rather than in discrete spaces because, while a finite model suffices for any particular example, no one finite model suffices for all examples.

The applications reported in this paper were selected to make the CIM model and its spatial representations more concrete to the reader, rather than seeking to demonstrate superior performance over other models. Two well-known experimental simulations that had employed vector representations were re-created using images as representations. The simulation of recognition memory in word list experiments closely followed assumptions made in implementing REM. Instead of the 20-dimensional vectors used as simulated words by Shiffrin & Steyvers (1997) we used images composed additively of between 15 and 20 randomly located circular Gaussians. The images used to represent perceived stimuli in Nosofsky's categorisation experiments were rudimentary, consisting of two non-intersecting subimages carrying information about values on the two experimental dimensions. We investigated representing magnitudes using circular Gaussians, complementary erf functions, and unstructured samples defined on grid approximations to the real plane.

Parameters in the simulations were not tuned to improve performance. This was because translating numerical vectors into images is really at cross purposes to the argument of this paper that spatial representations are a useful and biologically plausible alternative to representation using numbers. Nevertheless, the re-created simulations serve to demonstrate that the spatial representations and associated process model can explain cognitive experimental results, and are comparable with conventional modelling.

In addition to demonstrating the model, the simulations yielded an example of how CIM could explain an experimentally observed phenomenon, namely list strength effects, not explicable in the same way when feature values are modelled as numbers. This was because the explanation was due to memorising incomplete parts of feature values.

Further research into biologically inspired image representations of stimuli is obviously needed. Simulation is also needed of many more experiments, such as free and cued recall in memorisation experiments, or memory manipulation tasks.

The definition of working memory and the associated algorithm used for attention setting are not predicated on spatial or continuous representations. Our process model is inadequate as a general purpose cognitive system. For example, it was indicated that prototypes would be formed in some secondary phonological processing component but this was not described. The nature and interaction of multiple processing units needs to be investigated. The hard limit on the number of layers in working memory could be replaced by a soft limit, where a new layer created with each new probe is assumed to have a finite life, during which its total energy or activation subsides to zero. The practical importance of the size of working memory also needs to be examined, with its

effects on attention. Whether attention should decay gracefully, and whether circumstances exist in which it is rapidly cancelled, are amongst many modelling decisions that could be informed by biological parallels.

The operation of the CIM model on categorisation and recognition was roughly translated to the operation of a three layer neural net (Fig. 3), as an attempt to help readers understand it in terms of a familiar framework. This translation runs the danger of obscuring the important and novel characteristics of our modelling. For a start, the translation did away with working memory. Assigning a node for each memory may hide the fact that images are distributed representations of information, just as much as a representation learned by a neural net is distributed. In particular, fragments of a probe can be used for retrieval, and fragments can be combined. Our modelling should rather be seen as sitting above neural modelling, as the images are simulated records of the coherent activity of large numbers of neurons.

Learning in our model is accomplished simply by adding new memories. But what does it mean to add a memory? Memory is a pattern of activation that is in some sense recognisable as a unit. The *conjunctive neurons* described in Damasio's (1989) convergence zone theory are a mechanism for binding elements according to Simmons & Barsalou (2003). Such neurons bind states of feature maps, allowing associative mappings within modalities. Another set of conjunctive neurons bind across modalities. A conjunctive neuron, either intra- or cross-modality, could act as an index to an activation pattern. This would suffice to add that pattern as a memory to a store, although the actual mechanisms will undoubtedly prove more complex. In fact, Simmons & Barsalou (2003) extend Damasio's convergence zone theory with a so-called similarity-in-topography principle, which relates spatial proximity of conjunctive neurons to the similarity of the feature maps bound by the conjunctive neurons.

Eq. (15) set the stage for composition or binding with the internal generation of a probe, but this was not investigated. Prototypes in our simulations were generated from activated memories at the time they were required, and the index part of images was not explicitly simulated. Composition is the mechanism needed to create prototypes from exemplars, and to bind names or indices to them. It is a mechanism for extending semantic memories when new information about a concept is acquired. It is how the system combines concepts and is key to reflection and problem solving. An outstanding question is: under what circumstances would internal generation of a probe take precedence over an external interrupt? Binding/composition would obviously be sensitive to the number of layers in working memory, so this too needs investigation.

To go further with promoting CIM as a biological model, the set X on which the images are defined has to be described in terms of modality-specific areas of cortex. This will bring into question the assumption of a two dimensional patch: more general two-dimensional surfaces, or even three dimensional surfaces may be necessary. Greatly simplified cortical atlases should still be useful for predicting and explaining human cognition, just as connectionist models are useful despite being simplified.

Continuous spatial representations accompanied by a plausible process model have the potential to resolve many of the conundrums that have been arguably artificially raised by modelling psychological spaces inappropriately with vector representations. On the other hand, the choice of families of functions to use in modelling is ad hoc. There are a wealth of questions to be addressed in subsequent research, the answers to which will modify the model put forward here. However, the biological rationale for using continuous spatial functions of activity to represent information in cognition models will remain.

References

Abbott, L. F., & Dayan, P. (1999). The effect of correlated variability on the accuracy of a population code. *Neural Computation, 11*(1), 91–101.

Aisbett, J., & Gibbon, G. (2001). A general formulation of conceptual spaces. *Artificial Intelligence, 133*, 189–232.

Aisbett, J., & Gibbon, G. (2003). Preserving similarity in representation: a scheme based on images. In *Proceedings of*

the joint international conference on cognitive science, Sydney.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111*(4), 1036–1060.

Ashby, F., & Townsend, J. (1986). Varieties of perceptual independence. *Psychological Review, 93*(2), 154–179.

Baddeley, A. (1986). *Working memory*. Oxford: Clarendon Press.

Baddeley, A., & Logie (1999). Working memory: The multiple-component model. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 28–61). Cambridge University Press.

Barsalou, L. (1999). Perceptual symbol systems. *Behavioural and Brain Sciences, 22*, 577–660.

Blake, C., Keogh, E., & Merz, C. (1998). *UCI Repository of machine learning databases*. Irvine, CA: University of California, Dept. Information & Comp. Sci., Available from http://www.ics.uci.edu/~mlearn/MLRepository.html.

Courtney, S., Petit, L., Haxby, J., & Ungerleider, L. (1998). The role of prefrontal cortex in working memory: examining the contents of consciousness. *Philosophical Transactions of the Royal Society of London Series B, 353*(1377), 1819–1828.

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information processing system. *Psychological Bulletin, 104*, 163–191.

Damasio, A. R. (1989). Time locked multiregional retroactivation: a systems-level proposal for the neural substrates of recall and recognition. *Cognition, 33*, 25–62.

Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neuroscience, 23*, 475–483.

Eliasmith, C. (1996). The third contender: a critical examination of the dynamicist theory of cognition. *Philosophical Psychology, 9*(4), 441–463.

Engel, A. K., & Singer, W. (2001). Temporal binding and the neural correlates of sensory awareness. *Trends in Cognitive Sciences, 5*(1), 16–25.

Erwin, E., Obermayer, K., & Schulten, K. (1992). Formation of dimensional-reducing somatopographic maps. In S. Sayegh (Ed.), *Proceedings of the Fourth Conference on Neural Networks and Parallel Distributed Processing* (pp. 115–126). Indiana University at Fort Wayne.

Freeman, W. (1994). Qualitative overview of population neurodynamics. *Neural Modeling and Neural Networks*, 185–215.

Goldstone, R. (2003). Learning to perceive while perceiving to learn. In R. Kimchi et al. (Eds.), *Perceptual organisation in vision: behavioural and neural perspectives* (pp. 233–278). Lawrence Erlbaum Associates.

Haxby, J., Gobbini, M., Furey, M., Ishai, A., Schouten, J., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*, 2425–2430.

Kalish, M., & Kruschke, J. K. (2000). The role of attention shifts in the categorization of continuous dimensioned stimuli. *Psychological Research, 64*(2), 105–116.

Kosslyn, S. M. (2003). Understanding the mind's eye... and nose. *Nature Neuroscience, 6*(11), 1124–1125.

Kruschke, J. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological Review, 99*(1), 22–44.

Kruschke, J. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology, 45*, 812–863.

Laird, J., Newell, A., & Rosenbloom, P. (1987). Soar: an architecture for general intelligence. *Artificial Intelligence, 33*, 1–64.

Liley, D., Cadusch1, P., & Wright, J. (1999). A continuum theory of electro-cortical activity. *Neurocomputing, 26–27*, 795–800.

McFadden, J. (2002). Synchronous firing and its influence on the brain's electromagnetic field: evidence for an electromagnetic theory of consciousness. *Journal of Consciousness Studies, 9*(4), 23–50.

Miyake, A., & Shah, P. (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. New York: Cambridge University Press, pp. 442–481.

Norman, K. (2003). Computational models of Episodic Memory Art. 444. *Encyclopedia of cognitive science*.

Nosofsky, R. (1986). Attention, similarity and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115*(1), 39–57.

O'Reilly, R., Norman, K., & McClelland, J. (1998). A Hippocampal Model of recognition memory. In M. I. Jordan, et al. (Eds.). *Advances in neural information processing systems* (Vol. 10, pp. 73–79). MIT Press.

Poldrack, R. A. (2000). Imaging brain plasticity: conceptual and methodological issues – a theoretical review. *Neuroimage, 12*(1), 1–13.

Pylyshyn, Z. W. (2002). Mental imagery: in search of a theory. *Behavioral and Brain Sciences, 25*, 157–238.

Ratcliff, R., Clark & Shiffrin, R. (1990). List-strength effect. *Journal of Experimental Psychology: Learning, Memory & Cognition, 16*, 163–178.

Ryle, G. (1949). *The concept of mind*. University Chigago Press. Reprint edition 1984.

Saunders, B. A. C., & van Brakel, J. (1997). Are there nontrivial constraints on colour categorization? *Behavioral and Brain Sciences, 20*, 167–228.

Schyns, P., Goldstone, R., & Thibaut, J.-P. (1998). The development of features in object concepts. *Behavioural and Brain Science, 21*, 1–54.

Shepard, R. (1966). Metric structures in ordinal data. *Journal of Mathematical Psychology, 3*, 287–315.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM – Retrieving effectively from Memory. *Psychonomic Bulletin & Review, 4*(2), 145–166.

Shiffrin, R. M. (2003). Modeling memory and perception. *Cognitive Science, 27*, 341–378.

Simmons, W. K., & Barsalou, L. (2003). The similarity-in-topography principle reconciling theories of conceptual deficits. *Cognitive Neuropsychology, 20*, 451–486.

Sloman, A. (1978). *The Computer Revolution in Philosophy: Philosophy, Science And Models of Mind*. Harvester Press and Humanities Press.

Tijsseling, A., & Gluck, M. (2002). A connectionist approach to processing dimensional interaction. *Connection Science, 12*(1), 1–48.

Townsend, J., & Thomas, R. (1993). On the need for a general quantitative theory of pattern similarity. In S. C. Masin (Ed.), *Foundations of perceptual theory. Advances in psychology* (99, 297–368). North Holland.

Townsend, J. T., Solomon, B., & Spencer-Smith, J. (2001). The perfect Gestalt: infinite dimensional Riemannian face spaces and other aspects of face perception. In M. Wegner & J. Townsend (Eds.), *Computational geometric and process perspectives on facial cognition* (pp. 39–82). Erlbaum.

Townsend, J., & Spencer-Smith, J. (2003). Two kinds of global perceptual separability and curvature. In C. Kaernbach, E. Shroger, & H. Muller (Eds.), *Psychophysics beyond sensation: Laws and invariant of human cognition*. Erlbaum.

Treisman, A. (1999). Solutions to the binding problem: progress through controversy and convergence. *Neuron, 24*, 105–110.

Tulving, E., & Markowitsch, H. J. (1998). Episodic and declarative memory: role of the hippocampus. *Hippocampus, 8*, 198–204.

Van Gelder, T., & Port, R. (1995). It's about time: an overview of the dynamical approach to cognition. *Mind as motion: Explorations in the dynamics of cognition*. MIT Press.

Wang, X., Hutchinson, R., & Mitchell, T. (2003). Training fMRI classifiers to detect cognitive states across multiple human subjects. In *Proceedings of the neural information processing systems conference*.

Whalen, J., Gallistel, C., & Gelman, R. (1999). Nonverbal counting in humans: the psychophysics of number representation. *Psychological Science, 10*(2), 130–138.