Selective Attention and the Acquisition of Spatial Semantics

James M. Hogan, Joachim Diederich and Gerard D. Finn
NeuroComputing Research Centre, QUT,
GPO Box 2434, Brisbane, Q, 4001.
{hogan,joachim,gerry}@fit.qut.edu.au

Abstract

The acquisition of the semantics of natural language spatial terms is considered within the cognitive framework introduced by (Langacker, 1987), and the computational framework of the Berkeley L_0 project (Feldman et. al., 1990). We describe a computational model which incorporates selective attention mechanisms to facilitate the identification of significant objects within the visual field, and their consequent binding to linguistic relational identifiers (for example, the trajector and landmark) according to the conventions of the input language. In contrast to previous work in this area, the approach allows extension of the system to more sophisticated (potentially cluttered and feature-laden) input scenes and referential linguistic phenomena, without a major redesign of the system.

The application of the model to lexemes describing static concepts such as the English *above, below* and *in* is discussed, as are extensions to dynamic concepts.

1 Introduction

This paper is concerned with the acquisition of natural language spatial semantics by a neurally plausible connectionist system. Within the cognitive framework introduced by (Langacker, 1987), elementary spatial concepts (such as the English *above*) are characterised by locative relations between a potentially mobile object called the *trajector* (TR) and a static reference object called the *landmark* (LM). Previous computational investigations of this problem (Regier, 1992), have relied upon highly structured feature detection systems and the abstraction of object identification issues into the input data. While highly successful on their own terms, systems of this nature are not readily generalisable to problems involving more sophisticated (especially cluttered) input scenes and linguistic phenomena, and provide neither a conscious nor an autonomous selection mechanism through which such inputs may be successfully processed.

The model discussed below resolves some of these issues through the use of mechanisms of selective visual attention, through abstraction of established models from computational neuroscience (Niebur and Koch, 1997) and extension to allow linguistic input to cue selection and scene parsing. While retaining the overall computational philosophy of the Berkeley L_0 project (see section 2), the present work does not rely upon feature pre-processing to the same extent as the Regier system – representations being based upon probabilistic receptive fields. In this way, 'prior knowledge' of limited specificity may be employed through higher level recruitment to represent quite complex relations (see section 5.1 and (Hogan and Diederich, 1994), (Hogan and Diederich, 1995)).

The computational philosophy of the Berkeley L_0 project is introduced in the next section, followed by discussion of the Regier model and the importance of explicit object recognition in the light of evidence from early language acquisition. Section 2.4 relates this discussion to an accepted cognitive theory mediated through binding of representations at the focus of attention. Selective visual attention, and recent computational models of the process dominate chapter 3, prior to a formal outline of the model in chapter 4. The paper concludes with examination of representations for a limited set of English static concepts – developed through simulations based upon novel Gaussian domain response units – along with discussion of extensions to dynamic concepts.

2 Connectionist Modelling and the L_0 Project

Advances in brain sciences and information technology in recent decades have allowed the development of sophisticated models of cognitive processes at a

James M. Hogan, Joachim Diederich and Gerard D. Finn (1998) **Selective Attention and the Acquisition of Spatial Semantics.** In D.M.W. Powers (ed.) *NeMLaP3/CoNLL98: New Methods in Language Processing and Computational Natural Language Learning,* ACL, pp 235-244.

number of levels of abstraction. Noting the domain-specific nature of much of this work, and the importance of integration of disparate cognitive machinery in the long-term development of the discipline, (Feldman et. al., 1990) proposed L_0 as a "touchstone [task] for cognitive science", requiring elements of visual perception, natural language modelling, and learning. As originally stated, the L_0 task is to construct a computer system to perform Miniature Language Acquisition, without reliance upon "forthcoming results in related domains" to resuscitate an otherwise inadequate model:

> *The system is given examples of pictures paired with true statements about those pictures in an arbitrary natural language. The system is to learn the relevant portion of the language well enough so that given a novel sentence of that language, it can determine whether or not the sentence is true of the accompanying picture.*

The system is further constrained by the substantial variations known to exist across natural languages in their characterisation of space – eliminating *ad hoc* computational mechanisms – and by the assumption that learning must simulate childhood language acquisition in the exclusion of explicit negative evidence (see for example (Chomsky, 1965)). Thus only positive instances of a given concept may be presented during training, but the system may receive negative examples during normal operation.

2.1 A Semantic Sub-Task

The L_0 sub-task examined by (Regier, 1992) requires that the model system acquire "perceptually grounded semantics of natural language spatial terms". Each lexeme describes a locative relationship between a special (potentially mobile) object known as the trajector (TR) and a static reference object known as the landmark (LM) (Langacker, 1987). Figure 1 shows a positive example for the English lexeme 'above'. In essence, spatial semantics define a partitioning of the set of object pictures into classes prescribed by the underlying natural language. The task of the model system is then to learn this classification from positive examples of each category, forming a recognition system for each class of pictures[1]. In English, these labels might include the

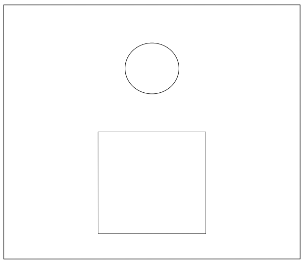

Figure 1: *Example image for the L_0 task, with which one might associate the English fragment "circle above square".*

static concepts: *above, below, left, right, in, out, on*, and *off;* and their dynamic equivalents: *above, below, left, right, around, in, on, out of, through*, and *over*.

System input is provided in the form of a two-dimensional bitmap (static concepts) or sequence of bitmaps (hereafter a "movie", for dynamic concepts) usually showing only the 'line-drawn' LM and TR in a position exemplifying the concept, although Regier does also consider more complicated phenomena such as *deixis*[2]. The task is thus simplified so as to limit issues of object detection (through avoidance of feature-laden scene backgrounds and object interiors), confusion due to distractors, and to disregard the role of luminance and colour.

Nevertheless, computational approaches are greatly constrained by many years of research in a number of disciplines, rendering the task of feature extraction and encoding non-trivial. While a cognitive model need not replicate all aspects of the underlying neural substrate, it gains in plausibility if it supports classifications based upon processing in cortical areas known to be active during performance of the given task. Thus, some functional replication of neural pathways – ostensibly at the level of systems neuroscience (Churchland and

[1]However, examples may belong to a number of categories, and some gradation of class membership is desirable as some scenes are better, more *prototypical* exemplars of a given concept than others. See chapter 2 of (Regier, 1992) for discussion of this issue.

[2]These extensions are not examined in the present work.

Sejnowksi, 1992) – becomes an essential aspect of architectural design, and this is more readily accomplished through a top down approach.

2.2 The Regier Model

(Regier, 1992) implemented highly structured connectionist systems for both the static and dynamic concept classes discussed above – the dynamic system incorporating the single frame processing capabilities of the static system. Concepts were represented in terms of *directional*[3] and *non-directional*[4] features computed from the image, system preprocessing providing the output network with real-valued encodings for each feature value. In contrast to the present model, objects are tagged as LM or TR tokens as part of the input representation, and the image is partitioned into separate TR and LM bitmaps as part of pre-processing.

Computationally, the Regier system may be viewed within the framework of "partially structured connectionism" (Feldman et. al., 1988), in which systems level architectural design is coupled with unstructured local networks which may be trained to perform (initially unspecified) functions so as to realise an overall system task – although this description understates the specificity of some model subsystems[5].

2.3 Discussion of the Regier Model

It is well-accepted that perceptual representations may rely upon independent encodings of object features and properties in distinct anatomical areas, and that some mechanism is then required to associate or *bind* the representations together to facilitate processing of a particular object instantiation (Treisman, 1996). This observation is best illustrated by the separation of the object recognition (variously the 'what'/'object'/'ventral'/'occipito-temporal') and location ('where'/'spatial'/'dorsal'/'occipito-parietal') pathways of the visual system.

While Regier carefully positioned his model clear of any controversy over correspondence with biological structures, its architecture must ultimately be viewed as an abstraction of the 'where' (dorsal) pathway, the need for object recognition being reduced through explicit tagging of the input data.

Although spatial relations are implicitly determined by the position of the objects in an example image, equally valid but semantically distinct (perhaps antonymic) characterisations of the scene may be made depending upon the selection of trajector and landmark. Figure 1, for example, may be regarded as prototypical example of both *above* ("Circle above square"; *TR=Circle; LM=Square*) or *below* ("Square below Circle"; *TR=Square; LM=Circle*). Identification of TR and LM is thus critical in the selection of the appropriate lexeme, and correct tagging appears to require association of an object name and internal representation sufficient to facilitate visual search, and a language-specific comprehension of the syntactic relationship between the TR,LM, and lexeme[6]. It is our contention that childhood acquisition of spatial semantics is dependent upon sufficient facility in the native language to perform this object-tagging, through parsing of spoken language fragments associated with the image.

It is thought (Crystal, 1995),(Khanji and Weist, 1996) that acquisition of elementary spatial lexemes takes place soon after the "naming explosion" of the second year of life (Woodward et. al., 1994), and prior to the development of sophisticated internal models of space (such as those allowing scene rotation and manipulation). Studies of spatial and temporal lexeme acquisition among young children native in various European and Middle Eastern languages (Johnston and Slobin, 1979), (Weist, 1991),(Khanji and Weist, 1996) indicate that subject groups of mean age as low as 30 months may correctly[7] associate pictures with spoken sentences such as "The parrot is in/on the cage".

[3] In particular, the orientation of a line connecting points of closest approach, and of that connecting the centres of mass.

[4] For example surface contact or inclusion.

[5] In the original L_0 paper, (Feldman et. al., 1990) noted the difficulty in balancing the facilitation of learning provided by "innate structures" (in computational modelling a top down approach) against the potential generality of relatively unstructured networks. Notwithstanding the apparent structural sophistication of the Regier model – perhaps motivated in part by difficulties in parameter adjustment with limited training sets – the choice of feature extraction machinery was in this case sufficiently general to allow lexeme acquisition across a variety of natural languages.

[6] Language-specificity to this extent does not violate the requirements of the L_0 task. While the Regier system was successfully applied to a number of natural languages, acquisition for a given language was performed independently of other training, utilising an output network encoding (and subtle adjustments to internal feature detector parameters) specific to that language. Syntactic variations are a similar limitation of system generality.

[7] Correctness is here a matter of statistically significant deviation from random performance, there being typically two alternative language fragments offered with a pair of images.

2.4 Relationship to Feature Binding

Issues of visual search and object recognition must necessarily assume greater importance with increases in the complexity of the scene – with consequent difficulty in tagging of TR and LM – but some linkage must be provided between object identification and location if spatially based semantics are to be encoded and processed. (Treisman, 1996), notes that object instantiation requires construction from more elementary features (such as shape and colour) and maintenance of the resulting entity through displacement or continuous motion. While the exact neural mechanisms which mediate binding are unknown, the most likely candidates are thought to involve temporary cell assemblies selected by focussed attention – with activations corresponding to the attended object remaining undiminished and those away from this region being suppressed. Propagation of these activations through a global location map provides the common reference point needed to link disparate representations (Treisman and Gelade, 1980).

Significantly, the cognitive framework discussed above was introduced to explain performance degradation of visual search within cluttered domains with complex feature conjunctions, and is closely aligned with the neural mechanisms considered in the following sections. A model based upon selective attention thus has the advantage of a unified approach to the disparate processing requirements of the problem, while providing a sound base for extensions to more complicated input scenes and linguistic phenomena.

3 Selective Visual Attention

It is well known that primate visual cortex receives information from the optic nerve at a rate well above the region's storage and processing capacities. Some mechanism of *selective attention*, whereby a small but important subset of the visual field may be given detailed processing is therefore necessary. Visual processing is typically decoupled into two regimes (Niebur and Koch, 1997):

- A pre-attentive phase, during which parallel extraction of elementary features is performed
- An attentive phase, during which the more salient or conspicuous stimuli within the field are processed in sequence, input from other stimuli being suppressed during this processing.

Attentional processing thus requires some selection mechanism based upon the elementary features extracted during the pre-attentive phase – with possible external input from some other neural region or sensory domain[8]. However, the selection mechanism need not be spatially sequential, and two types of *covert*[9] visual attention are commonly distinguished, governed largely by the nature of the (perhaps automatic) search task being undertaken by the visual system. *Focal attention* (Niebur and Koch, 1995),(Niebur and Koch, 1997) is a sequential search through a series of progressively less salient locations, selection being driven primarily from *below* – saliency being determined from the contributions of elementary features extracted during the pre-attentive phase. In contrast, *dispersed* or *feature-based attention* (Usher and Niebur, 1996), (Niebur and Koch, 1997) is spatially parallel, but regarded as sequential within some feature space – the selection relying upon some "top-down" signal to highlight a particular conjunction of features.

3.1 Neural Gating – The Saliency Map

Regardless of whether saliency is an emergent property of the input scene or imposed (perhaps consciously) from some other cortical region, each model requires that selection and suppression of stimuli be realisable in a neurally plausible structure. Most location-based attentional models are at present based upon the *saliency map*, introduced by (Koch and Ullman, 1985). While no localised neural implementation of this structure has been discovered, there is strong evidence for the existence of a mechanism based upon several elementary features extracted from the image (Niebur and Koch, 1997), and unit activations within the map are computed from a weighted sum of feature map outputs – giving a measure of "conspicuity" within each unit's receptive field[10].

(Niebur and Koch, 1995) employ a total of eight input maps based upon orientation, intensity, chromatic components and temporal change – along with provision for "external" (i.e. top-down) inputs to

[8] (Koch and Ullman, 1985) suggest that attentional control may be located as peripherally as the LGN, relying upon back-projections from cortical feature maps.

[9] High resolution visual processing is dependent upon alignment of fovea and stimulus, normally achieved in primates through rapid eye and head movements in a process known as *overt attention* (Niebur and Koch, 1997). Neither mechanism is considered in this brief review, and our model assumes that covert attentional shifts are sufficient to capture phenomena of interest – a simplification which must break down for wide field moving trajectors but is otherwise plausible.

[10] Note the similarity to the *master feature map* of Feature Integration Theory (Treisman and Gelade, 1980).

account for cueing effects. The most salient feature in the input field is then computed by means of a winner-take-all network over the map, selecting the unit with the highest activation and suppressing output from the remaining units through recurrent connections. In addition, the winning unit is itself inhibited over time, allowing attention to shift to a salient (but previously unattended) stimulus even if the scene remains unchanged. This inhibition serves also to prevent the immediate return of attention to a previously attended site, in accordance with psychophysical evidence (Posner and Cohen, 1984),(Tipper et. al., 1991).

3.2 Neural Gating – Object Representation

(Fujita et. al., 1992) found through cell recordings that neurons within infero-temporal (IT) cortex are organised into columns with optimal selectivity toward abstractions of known objects (simple geometric shapes, differential shading etc.) with activation significantly greater when presented with the abstracted or minimalist image rather than a detailed photograph of a similar object. On anatomical (i.e. resource limitation) grounds, these findings suggest that objects may be represented through a combination of no more than 1000 of these elemental pictures, with adaptation of representations occurring as necessary[11].

Usher and Niebur's model of feature-based attention (Usher and Niebur, 1996) receives input from the entire visual field through such activated IT cortex cell assemblies, with the search task guided by weak "top-down" activation of the favoured feature class from a similar representation in working memory (here taken to be pre-frontal cortex). While an explicit saliency map is not employed, the attended stimulus is again determined through competitive selection among the input representations (here object cell assemblies). In a cluttered field, top-down activation may provide a winning advantage to the favoured object.

3.3 Modulation at the Focus of Attention

Once the most salient stimulus has been selected from among its competitors, some mechanism must be employed to facilitate passage of its associated input data to "higher" cortical centres while suppressing passage of competing input. In the Niebur and Koch model (Niebur and Koch, 1995), a modulating signal from the saliency map is propagated via recurrent connections back to the region of primary visual cortex (V1) associated with the winning unit. Enhanced activation is thus re-propagated along the visual pathways, giving this input stream substantial advantages in any competitive selection processes subsequently encountered[12]. Widespread propagation of an enhancement signal of this kind to features associated with an object at the most salient location in the visual field is thought to underpin feature binding (Treisman, 1996).

[11] Note that these columns have low spatial selectivity – existing well along this visual processing hierarchy – and are sensitive to such stimuli regardless of their position in the field.

[12] Enhancement of activation is accomplished via *temporal tagging* – modulation of the spike train through a time-varying Poisson process (Niebur and Koch, 1995).

4 Model Architecture

This section introduces a connectionist model for spatial lexeme acquisition based upon the attentional mechanisms discussed above. Only the model for static concepts is presented here, although few changes are necessary to the gross architecture to accommodate the dynamic case. As in the Regier model, an unstructured output or decision network encodes the lexeme representations, receiving input from neurally inspired processing modules – although here the object recognition pathway is explicitly considered. The following sections outline the gross architecture and functionality of the model, developing each substructure in turn before discussion of the output network. Implementation and representation issues are examined in section 5.

4.1 A Conceptual Model for the Static Case

Each static scene may be characterised as a movie consisting of repeated presentations of the same frame – attention initially focussed upon one object (for example the TR) and passing during movie presentation to the other (the LM). Network learning depends upon presentation of frames exemplifying each of these phases, and object tagging (identification of objects as respectively TR and LM) relies upon "visual search" initiated by parsing of the language fragment, and subsequent binding of object feature and location information. The approach is solidly grounded in the Feature Integration Theory of Treisman (Treisman and Gelade, 1980), with perceptual binding mediated through selective attention.

4.2 The Recognition Pathway

Processing corresponding to the early visual system is not explicitly modelled, and system input is provided by three unit banks, representing language input, object recognition and object location. Object representation is based upon the IT cortex assemblies discussed in section 3.2, with the simplifying assumption that input scenes contain only objects closely identifiable with a single iconic image – the system being restricted to a discrete set of object types whose presence is indicated by the activation of a single input unit[13].

Language input is similarly reduced to a bank of object units, on the basis that apprehension of the object description (for example a simple noun such as *circle*) is sufficient to activate a representation of the object, already available in memory as a result of exposure to the image. In computational terms, the visual object has been tagged as a `CIRCLE` token, and the iconic `CIRCLE` representation activated, although the reality is far less neatly partitioned. This representation provides top-down activation in much the same manner as the working memory module of the Usher and Niebur model (Usher and Niebur, 1996), the mechanisms together realising object tagging through an abstraction of feature-based visual search.

The relationship between language and object input is shown at the right of figure 2, tagging being represented by a conjunction between the language and object units within the binding network – the winning conjunction being selected through a Winner-Take-All (WTA) network (Feldman, 1982), and unwanted, weaker conjunctions being discarded. Such selection and suppression mechanisms readily allow generalisation of the tagging system to more cluttered scenes or sophisticated linguistic phenomena, particularly as tagging is performed over time – greatly reducing problems of cross-talk.

The robustness of the cell assembly representation is here captured through multiple random projections from each unit to the binding network, ensuring with high probability that *at least one* connection with a particular binding unit is realised[14].

The function of the binding subsystem is illustrated in the following table by the example of of figure 1 and the language fragment "circle above square". Input from the object assemblies remains constant throughout the period, and for the sake of brevity is suppressed. For clarity, the number of scene frames is limited to four, with change in the language input after the second frame:

[13]Extensions to more complicated objects require representation in terms of a weighted combination of these iconic 'letters'. The binding mechanisms discussed here are in principle sufficient to handle such extensions, but pre-processing would necessarily be complex.

[14]These reliability arguments are based upon those advanced by Feldman (Feldman, 1982).

Frame	Lang Circle	Lang Square	Binding
0	1	0	$TR < Circle >$
1	1	0	$TR < Circle >$
2	0	1	$LM < Square >$
3	0	1	$LM < Square >$

4.3 Integration with the Location Pathway

For object tagging to be useful in the present context requires some integration of the feature and location based models of selective attention considered in section 3. The mechanisms of the previous section are strongly reliant upon feature-based attention (Usher and Niebur, 1996), and do not require an explicit saliency map.

Recall that location-based attentional models (Niebur and Koch, 1995), construct saliency as a weighted sum of several constituent feature maps – which while representing anatomically distinct areas, provide inherent location binding. The model also provides for external input to this map to account for cueing – perhaps mediated through representations in working memory – but again the input is location bound.

The current work preserves the global saliency map of (Niebur and Koch, 1995), but introduces feature based input to the map through the external channel of the previous paragraph, as though the primitive object cell assemblies of IT cortex were merely another feature map contributing to overall saliency. Both classes of model (colloquially 'where' and 'what') rely on top-down modulation of activation in order to implement the selection of the attended region. In the former case, modulation takes place through recurrent connections to primary visual cortex, and 'where' to 'what' information transfer may take place through binding at the focus of attention – essentially through lock-step repropagation of the modulation along both pathways – although this is not required for the present task. 'What'-to-'where' transfer in the current model is based upon an extension of the feature-based model of (Usher and Niebur, 1996), with propagation of top-down modulation from the IT assemblies to striate cortex, and re-propagation as for the 'where'-'what' linkages[15].

[15]Only limited success has been achieved to date in elucidating mechanisms of communication between the two pathways, although binding of representations necessarily demands it. The model proposed here is attractive and plausible, but remains to be established experimentally (Niebur, 1997).

This mechanism is abstracted in the current model so that object-based input is effectively represented in another feature map, although some delay to account for the traversal of the pathway may be desirable in more sophisticated extensions. However, the approach effectively eliminates the need for direct input from the object assemblies to the decision network, as binding has been extended to the saliency map. As before, we may characterise this interaction by examining the bindings realised. The input sequence is as before, but suppressed for clarity, and location input is restricted to representative vectors x_1 (the square) and x_2 (the circle).

	Binding Network	Object Location Input	Implicit SM Binding
0	$TR < Circle >$	(x_1, x_2)	$TR(x_1)$
1	$TR < Circle >$	(x_1, x_2)	$TR(x_1)$
2	$LM < Square >$	(x_1, x_2)	$LM(x_2)$
3	$LM < Square >$	(x_1, x_2)	$LM(x_2)$

4.4 Lexeme Binding

As in the Regier model, lexeme acquisition is ultimately accomplished through a sparse-coded representation at an unstructured (here randomly connected) output network. In its purest form, the model exerts very tight control over the information which is passed to this decision network – object and location information being effectively gated by the saliency map. This decoupling of the problem both simplifies and complicates the issue: binding at the output network requires a lower degree conjunction, but the lexeme is now in principle a temporal rather than spatial conjunction – necessitating a recurrent output network.

Bottom up saliency is of relatively little consequence in the static case, as the conscious selection implied by the object tagging mechanism controls the focus of attention, and these considerations cannot be over-ridden by an unchanging input scene – although decay of the most salient location helps facilitate the attention shift.

Figure 2 shows the gross architecture in its entirety. Lexemes are represented by individual output units of the decision network, gated input being provided to this network from the saliency map, and language input (i.e. encoding of the lexeme itself) implicit in the learning mechanism. At this point, the network must represent a binding of the form:

$$above < TR(x_1), LM(x_2) > .$$

Successful acquisition of such bindings is dependent upon the structure of the saliency map and its relationship to the output network, and these issues are considered in detail in the following sections.

5 Representation and Model Implementation

As envisaged by Koch and successive co-authors, it is the role of the saliency map to determine the most salient input region, and to gate visual input so as to highlight this attended region for more detailed processing. In this way, selective attention is sited conceptually amongst detection of elementary features, and decoupled from more sophisticated representations computed further along the visual pathway. Yet while the selection mechanism typically isolates a salient region for high-resolution representation and processing (in part accomplished through suppression of competing stimuli), modulation may also be reflected in a relatively low-resolution representation of the entire field – highlighting the attended object at the expense of less salient ones. As the modulating signal is thought to be directed back to primary visual cortex, such *reduced maps* may be computed at a number of points along the visual processing hierarchy, as required by the sophistication of the relation to be represented[16].

While acknowledging, therefore, the importance of pre-processing as identified by (Regier, 1992), the present work does not employ feature extraction machinery of the same sophistication. In part, this may be justified by noting that much of the computational difficulty of the problem is removed once the fovea has been positioned – limiting the class of examples with which the system may be faced. Yet a more powerful justification is philosophical: the representations considered below require neither a high level of genetic determinism nor a long period of inductive learning to become established.

5.1 Random Receptive Fields

(Hogan and Diederich, 1994), (Hogan and Diederich, 1995) considered a novel class of connectionist networks in which connectivity is determined randomly, in accordance with biologically plausible probabilities. Briefly, probability of connection between each pair of neurons is dependent upon the "distance" between them – the local probability α being constant within some local radius R of each

[16] Notwithstanding the clear separation in this discussion between the saliency mechanism and further processing, the reduced representations discussed below are computed directly from the saliency map on the grounds of computational simplicity.

unit, and decaying exponentially outside this region. This earlier work established that networks of moderate size may harbour small subnetworks (known as candidate architectures) which could be usefully *recruited* in the representation of Boolean concepts[17].

In the present work, the approach is extended to produce random receptive field units, receiving projections from a high-resolution input map (30×30 units) under similar probability and radius restrictions as those above. Reduced maps of this type provide a kind of probabilistic localisation – proximal objects being represented with high probability, and distal objects being represented with low (but still significant) probability. Thus, the representation is flexible within the bounds provided by foveal alignment, allowing significant fault tolerance in the boundary of each field. Computationally, the approach provides a substantial reduction of dimension, producing an encoding of the problem allowing recruitment at the output network without propagation of an error signal to the underlying representation.

We conclude this section with an example reduced map, demonstrating that simple receptive fields of this type are sufficient to discriminate concepts such as *above*[18]. In the limit of a large number of possible projections, the response of each *receptive field* unit may be modelled through the use of *Gaussian domain response units* [19] developed for this purpose and trainable through gradient-descent. The significance of these simulations therefore lies not in the method of acquisition of the representation, but rather in the fact that such a representation may perform successfully.

Figure 3 shows the combined (weighted) response map of receptive fields for *above* obtained by training on example images showing a smaller object above a larger object[20]. As will be clear from the graphic, the strong positive response to activation in the centre of the upper region ensures that the map provides strong identification of prototypical positive and negative examples. However, the decaying rather than hard-limiting response of the fields provides sufficient flexibility that weakly positive examples with typical locations and prototypical examples with atypical locations are also correctly identified. For example, table 5.1 shows results using this field, outputs encoded in the interval $[0, 1]$, with 1.0 indicating a strongly positive result. The approximate location of the TR with respect to the LM is indicated using points of the compass, and the weak positive examples were not assigned a numeric target value.

TR Position	True Value	Network Value
N	1.0	0.966
E	0.0	0.066
W	0.0	0.088
S	0.0	0.064
NE	N/A	0.357
NE	N/A	0.302
NW	N/A	0.579

6 Conclusions

In this work, we have developed a powerful new architecture for modelling the acquisition of spatial semantics, providing a number of advantages over previous approaches – in particular in its potential for application to more cluttered input scenes and linguistically complex phenomena. While discussion has centred upon a system which caters for static concepts, the system is immediately extensible to the case of dynamic concepts through the addition of a temporal change map to the model input (Niebur and Koch, 1995).

Representations introduced by the model are based on simple, probabilistic receptive fields encoding activation of the saliency map, and requiring limited prior knowledge and learning to be realised – having also substantial advantages in fault tolerance. In forthcoming work we shall present results for system learning from a wide range of static and dynamic concepts and examine extensions of the model to include linguistic description of faces based upon the spatial relationship between constituent features (for example, the shape and relative positions of nose, mouth and eyes).

Acknowledgements

The authors would like to acknowledge helpful discussions with Ernst Niebur and Claudia Brugman.

[17] This approach is based upon evidence from cognitive neuroscience – see (Ramachandran, 1993) for a review.

[18] Similar representations have also been obtained for other English directional concepts such as *below, left* and *right*. A representation specific to *in* cannot be demonstrated in this way, requiring the attentional mechanism to highlight a detectable change of state within the local region – the change only appearing over time.

[19] The unit response to the intensity of each input is weighted according to a Gaussian function of the distance between the input and the unit centre.

[20] Typical training sets include strongly positive examples, coupled with a similar number of strongly negative examples of the concept, randomly positioned and labelled manually. The network successfully generalises to unseen weakly positive examples.

References

Chomsky, N. 1965. Aspects of the Theory of Syntax. Cambridge, MA., MIT Press.

Churchland, P. and Sejnowski, T.J. 1992. "The Computational Brain". Cambridge, MA., MIT Press.

Crystal, D. 1995. The Cambridge Encyclopaedia of Language. Cambridge, UK., Cambridge University Press.

Feldman, J.A. 1982 Dynamic Connections in Neural Networks. *Biol. Cybernetics,* 46, pp27-39.

Feldman, J.A., Fanty, M. and Goddard, N. 1988. Computing with Structured Neural Networks. *IEEE Computer*, **21**, pp.91-104.

Feldman, J.A., Lakoff, G., Stolcke, A. and Hollbach Weber, S. 1990. Miniature Language Acquisition: A Touchstone for Cognitive Science. **TR-90-009**, *International Computer Science Institute.*

Fujita, I., Tanaka, K., Ito, M. and Cheng, K. 1992. Columns for Visual Features of Objects in Monkey Inferotemporal Cortex. *Nature*, 360, pp. 343-346.

Hogan, J.M. and Diederich, J. 1994. Random Neural Networks of Biologically Plausible Connectivity. *Proc. 2nd Australian Complex Systems Conference*, Rockhampton, September.

Hogan, J.M. and Diederich, J. 1995. Feasibility of Incremental Learning in Biologically Plausible Networks. *Proc. 6th Australian Conference on Neural Networks*, Sydney, February.

Johnston, J.R. and Slobin, D.I. 1979. The Development of Locative Expressions in English, Italian, Serbo-Croatian and Turkish. *Journal of Child Language*, **6**, pp.529-545.

Johnston, J.R. 1988. Children's Verbal Representations of Spatial Location. In: Stiles-Davies, J. et. al. (Eds), *Spatial Cognition.* Hillsdale, NJ., Lawrence Erlbaum, pp.195-205.

Khanji, R. and Weist, R.M. 1996. Spatial and Temporal Locations in Child Jordanian Arabic. *Perceptual and Motor Skills*, **82**, pp.675-682.

Koch, C. and Ullman, S. 1985. Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry. Human Neurobiology, 4, pp.219-227.

Langacker, R. 1987 Foundations of Cognitive Grammar I: Theoretical Prerequisites. Stanford, CA., Stanford University Press.

Niebur, E., and Koch, C. 1995. Control of Selective Visual Attention: Modeling the "Where" pathway. Proceedings NIPS*95.

Niebur, E., and Koch, C. 1997. Computational Architectures for Attention. In: Parasuraman, (Ed)., *The Attentive Brain.* Cambridge, MA., MIT Press.

Niebur, E. 1997. Personal Communication.

Posner, M.I. and Cohen, Y.A. 1984. Components of Visual Orienting. In: Bouma, H. and Bouwhuis, D.G. (Eds), Attention and Performance X. Hillsdale NJ. Lawrence Erlbaum and Associates pp.531-554.

Ramachandran, V.S. 1993. Behavioral and Magnetoencephalographic Correlates of Plasticity in the Adult Human Brain. *Proc. Natl. Acad. Sci. USA,* **90**, pp. 10413–10420.

Regier, T. 1992. The Acquisition of Lexical Semantics for Spatial terms: A Connectionist Model of Perceptual Categorization. PhD Dissertation, University of California, Berkeley.

Tipper, S.P., Driver, J. and Weaver, B. 1991. Object Centred Inhibition of Return of Visual Attention. *Q.J. Experimental Psychology*, **43A**, pp.289-298.

Treisman, A. and Gelade, G. 1980. A Feature Integration Theory of Attention. *Cognitive Psychology*, **12**, pp. 97-136.

Treisman, A. 1996. The Binding Problem. *Current Opinion in Neurobiology*, **6**, pp. 171-178.

Usher, M. and Niebur, E. 1996. A neural model for parallel, expectation-driven attention for objects. *J. Cognitive Neuroscience*, **8**, pp.311-327.

Weist, R.M. 1991. Spatial and Temporal Location in Child Language. *First Language*, **11**, pp.253-267.

Woodward, A.L., Markman, E.M., and Fitzsimmons, C.M. 1994. Rapid Word Learning in 13- and 18-Month-Olds. Developmental Psychology, **30**, pp.553-566.

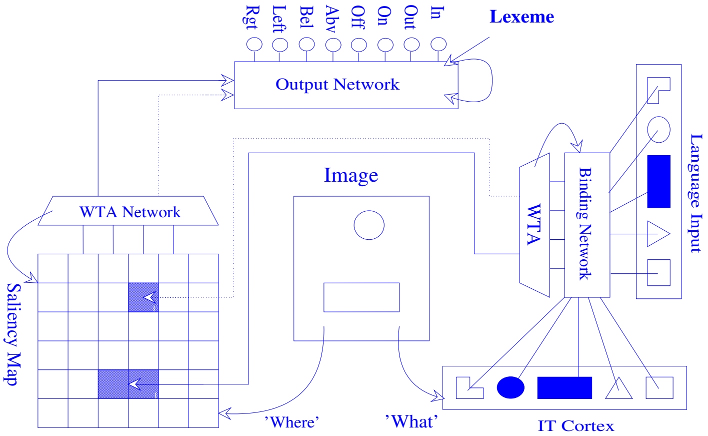

Figure 2: *The complete system architecture.*

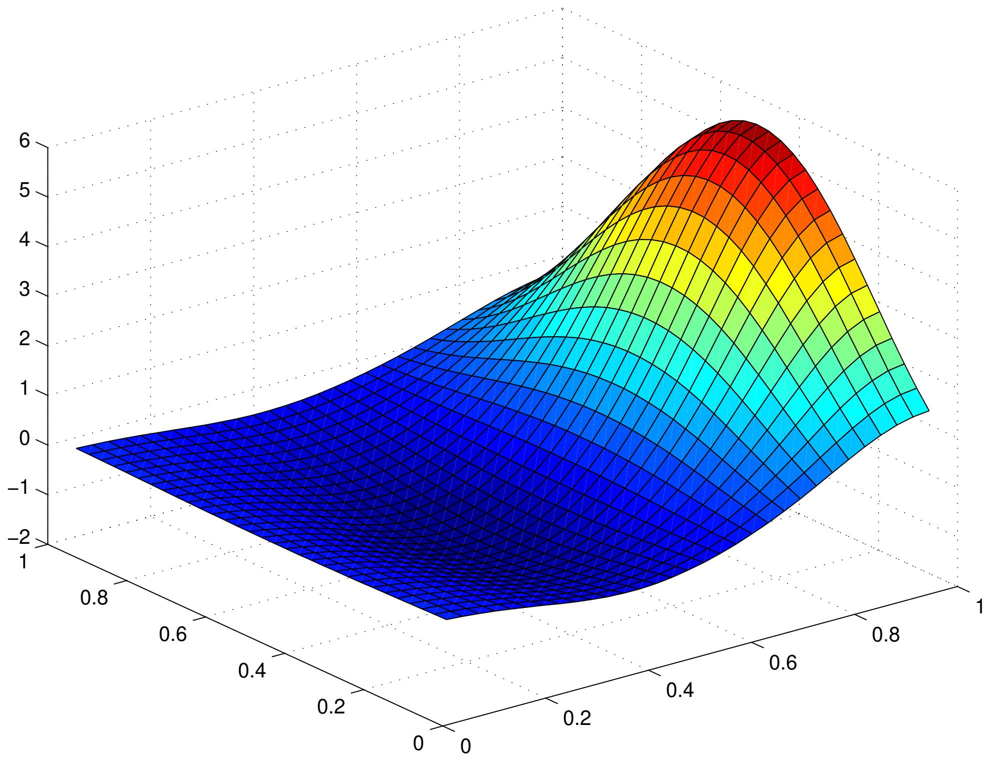

Figure 3: *Example receptive field map for* above.