Connection Science
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713411269

A PDP model of the simultaneous perception of multiple objects

Cynthia M. Henderson[a]; James L. McClelland[a]
[a] Department of Psychology, Stanford University, Stanford, CA, USA

Online publication date: 27 May 2011

To cite this Article Henderson, Cynthia M. and McClelland, James L.(2011) 'A PDP model of the simultaneous perception of multiple objects', Connection Science, 23: 2, 161 — 172

To link to this Article: DOI: 10.1080/09540091.2011.575931

URL: http://dx.doi.org/10.1080/09540091.2011.575931

Connection Science
Vol. 23, No. 2, June 2011, 161–172

A PDP model of the simultaneous perception of multiple objects

Cynthia M. Henderson* and James L. McClelland

Department of Psychology, Stanford University, Building 420, 450 Serra Mall, Stanford, CA 94305, USA

(*Received 14 March 2011; final version received 16 March 2011*)

Illusory conjunctions in normal and simultanagnosic subjects are two instances where the visual features of multiple objects are incorrectly 'bound' together. A connectionist model explores how multiple objects could be perceived correctly in normal subjects given sufficient time, but could give rise to illusory conjunctions with damage or time pressure. In this model, perception of two objects benefits from lateral connections between hidden layers modelling aspects of the ventral and dorsal visual pathways. As with simultanagnosia, simulations of dorsal lesions impair multi-object recognition. In contrast, a large ventral lesion has minimal effect on dorsal functioning, akin to dissociations between simple object manipulation (retained in visual form agnosia and semantic dementia) and object discrimination (impaired in these disorders) [Hodges, J.R., Bozeat, S., Lambon Ralph, M.A., Patterson, K., and Spatt, J. (2000), 'The Role of Conceptual Knowledge: Evidence from Semantic Dementia', *Brain*, 123, 1913–1925; Milner, A.D., and Goodale, M.A. (2006), *The Visual Brain in Action* (2nd ed.), New York: Oxford]. It is hoped that the functioning of this model might suggest potential processes underlying dorsal and ventral contributions to the correct perception of multiple objects.

Keywords: simultanagnosia; illusory conjunctions; object recognition; connectionism; feature binding

1. Introduction

Lesion studies have helped to advance the view that visual processing is largely split into 'what' and 'where'/'how' pathways in the temporal and parietal cortices, respectively. The famous case of D.F. illustrated that basic object manipulations, such as grasping, were still possible even after neural damage severely impaired her ability to recognise the objects she grasped visually. Other studies have found that macaques and humans with parietal lesions can retain their ability to discriminate objects, even if visually guided grasping or localising of objects can be plagued with mistakes (Milner and Goodale 2006).

However, curiously, profound failures in the perception of *multiple* objects occur after bilateral posterior parietal lesions, in a condition called simultanagnosia or Balint's syndrome. Sufferers of this disorder experience difficulty in such everyday tasks as navigating through their households, reading sentences (citing confusion from multiple words), or perceiving more than one participant in a conversation (Coslett and Saffran 1991). Upon presentation of two similar stimuli,

*Corresponding author. Email: chenders@stanford.edu

ISSN 0954-0091 print/ISSN 1360-0494 online

DOI: 10.1080/09540091.2011.575931
http://www.informaworld.com

simultanagnosics might report seeing only one stimulus, but the reported stimulus might have some of the features of the other, non-reported (and presumably non-perceived) stimulus (Robertson, Treisman, Friedman-Hill and Grabowecky 1997). This phenomenon of erroneously combining the features of multiple stimuli has been named illusory conjunctions, and has been taken as evidence that visual perception includes a process wherein features of different objects become perceived as part of specific objects or at specific locations (Treisman 1998). In this paper, we focus on one sub-part of the binding problem, namely, how the features of two objects could be correctly combined or 'bound' to form a perception of their respective objects.

Conditions exist in which normal subjects also produce substantial illusory conjunctions. Experiments by Treisman and Schmidt (1982) and Mozer (1983) examined illusory conjunctions under situations of high attentional loads or brief presentation durations, respectively. In both of these studies, normal subjects routinely and erroneously perceived one stimulus as incorporating features of the other – in the case of Treisman and Schmidt, stimuli were colored letters or shapes, and in the case of Mozer, they were four-letter words. Importantly, both sets of experiments involved pairs of similar stimuli presented concurrently.

From the perspective of neural networks, the perception of multiple objects is difficult, in part because of the constraints imposed by needing to generalise learning across space. Consider, for example, a visual system which has learned to recognise a word such as 'FORK' in one portion of visual space. Such a system should also be able to recognise 'FORK' when it is presented elsewhere, or there would be a massive and unfortunate replication of learning as the visual system learned about every word at every point in space. One solution might be to simply encode the letters and their relative positions. However, to the degree that the only locational information encoded for a word is the relative positions of its letters, the system might have difficulty perceiving multiple objects concurrently. For example, attempts to concurrently process two words such as 'FORK' and 'POLE' might run into interference from alternative options with the same relative positions of the constituent letters, e.g. 'FORE', 'FOLK', 'PORK', or 'PORE'. While behavioural results support the existence of some types of recognition that are tolerant to changes in an object's retinal location (Remus 2010), the way that such a process could occur and yet allow correct perception of multiple objects in parallel is not fully understood. In the following section, we present an initial exploration of a neural network model that exhibits both the correct concurrent perception of two objects, as well as the erroneous featural 'binding' of illusory conjunctions found under brief viewing conditions and simultanagnosia.

Our network consists of two sections (Figure 1). A section designed to acquire similar functions to the dorsal visual pathway (hereafter called the dorsal section) transforms the inputs into a topographic map of distributed object-appropriate 'actions' that might be taken. In modelling the dorsal section in this manner, we subscribe to the Milner and Goodale (2006) view of the dorsal visual pathway's role in planning actions. Our dorsal section also has the emergent property of representing object features, in line with recent neuroimaging findings regarding the intraparietal cortex (Konen and Kastner 2008). Second, our model contains a 'ventral' section, with an output of spatially invariant localist object and category units. We find that the hidden layer sizes in the ventral section regulate a trade-off between processing information about multiple objects (improved with larger hidden layers) versus the emergent property of spatially invariant recognition of objects (improved with smaller hidden layers to compress the representations).

While both the dorsal and ventral components of our model generalise learning across space, they do so in distinct ways and so have different capabilities. The ventral representation of objects is largely spatially invariant in its second hidden layer; as a result, this system can learn idiosyncratic information about particular objects and generalise this across positions, but, when not connected to its dorsal counterpart, it often makes errors in recognising multiple items in parallel.

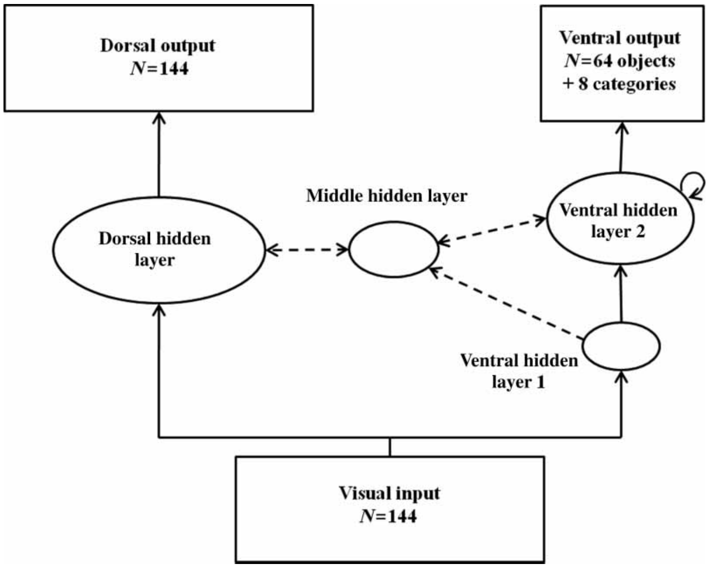

Figure 1. Schematic of the Dorsal–Ventral model. Rectangles indicate input and output layers; ovals indicate hidden layers. In the full model, dorsal and ventral hidden layers were connected through a middle hidden layer. The Dorsal–Ventral network whose results are reported here had hidden layer sizes of $N = 144$ for the dorsal hidden layer, $N = 50$ for the middle hidden layer, $N = 22$ for the ventral hidden layer 1, and $N = 65$ for the ventral hidden layer 2.

In particular, illusory conjunctions can occur, as ventral representations for the two similar presented objects activate representations for non-presented objects which share features with the two objects present.

In contrast to the ventral section of the model, the dorsal section creates distinct hidden layer representations when a given object is presented at different spatial positions. The dorsal section can respond correctly to multiple objects presented simultaneously and can even respond correctly to novel objects (i.e. novel combinations of features), but by itself it does not capture idiosyncratic information. By combining these sections via lateral connections, we create a network that can overcome most of the limitations of the ventral and dorsal sections and generalise idiosyncratic identity and category affiliations of objects to novel locations as well as perceiving in parallel the identity of more than one object at a time.

2. Methodology

The network simulations presented here provided us with a number of free parameters that we could adjust to modify their behaviour. In selecting values for these parameters, our consideration was for the network's ability to generalise across space (as described below). In the rare instance where we found more than one possible network with this ability, we chose the simpler one. No further changes to the network parameters were made to account for any of the experimental data.

As the weights in our networks were learned, the representations in the hidden layers were distributed. However, the hidden layers for each pathway were trained to produce either localist or distributed outputs and, as described below, characteristics of the outputs related to this distinction led to important differences between the two sections of the network.

2.1. *Inputs and outputs*

Inputs and target output patterns were abstract representations that bear no necessary relation to actual visual inputs or to the detailed properties of ventral or dorsal outputs. However, they were created to have some similar visual processing requirements. The decision to initially use abstract stimuli was made in the hope that their simplicity might lead to both a more parsimonious model and to more easily interpreted results.

2.2. *Input*

The ventral, dorsal, and combined networks all used the same input patterns. This input provided both spatial and featural information about the objects present. The input consisted of six options for features that were arrayed across 24 positions. The features are labelled as texture (plastic or metal), size (small or large), and rotundity (round or flat). A given 'object' could appear anywhere along the x-axis of the visual input and would occupy two positional spaces. As illustrated in (Figure 2: bottom), an object might be metallic, large, and flat on its left half, and metallic, large, and round on its right half (e.g. a large metallic spoon).

2.3. *Dorsal output*

The dorsal section is intended to model in a simplified form the process by which visual shape information might guide our interaction with objects. The output can be thought of as locations where one should optimally place one's hands to manipulate the objects. Each action in this system was unique for a given object: the material, size, and rotundity of both halves of an object determined the vertical displacement, horizontal displacement, and spread of the corresponding action pattern, respectively. For example, if an object was large on both sides, then it might be more optimal to position your hands more widely. Likewise, heavier, metallic objects might be better grasped by reaching more for the base of the object; and a round object might encourage a more focal point of contact than a flat object (e.g. see Figure 2: top). Knowledge of a single feature of an object was insufficient to determine which dorsal output units should be active, as the values of the other features of that object could completely change which output units should be active, and similar objects did not always overlap in their target output units. However, as was possible due to its distributed representation, each action was also systematically related to the actions for other, similar objects.

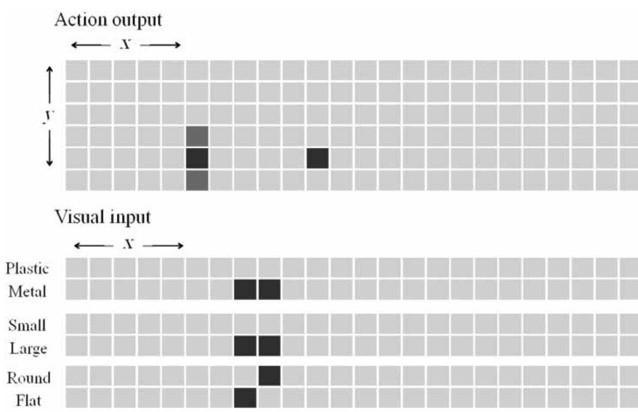

Figure 2. Visual input and output action for an object presented at the eighth position.

2.4. *Ventral output*

The ventral output consisted of localist units, representing the identity and category membership of the object presented to the network. There were 64 possible objects. A given object and a given category were represented by a single output unit each. Each object was randomly chosen to belong to one of eight possible categories; these categories were idiosyncratic properties of an object in that they were unrelated to the visual features of each object. The ventral output conveyed information about the identity and category of an object, but no information about its spatial position.

Given the localist outputs, it was not possible to have a systematic relationship between outputs for different objects; the identity outputs for different objects were thus unrelated and idiosyncratic.

2.5. *Processing and training details*

Units in these networks used a sigmoid activation function. The networks were trained with backpropagation and, with the exception of the feed-forward dorsal network, all networks were recurrent. Networks were trained using the PDPTool simulation package implemented in MATLAB (available at http://www.stanford.edu/group/pdplab/resources.html) using the bp (dorsal network) and rbp programs. Recurrent backpropagation simulations lasted either seven intervals (ventral network) or 10 intervals (combined network), using a time step size (dt) of 0.25 intervals. Networks used cross-entropy error; momentum was set to 0.9, learning rate was either 0.0005 (dorsal network) or 0.00005, and there was no weight decay.

We trained the networks fully (i.e. for every position) for 62 out of 64 objects; one object was selected to be completely novel, and one object was selected to be positionally restricted (#51). This positionally restricted object, #51, was only presented in the first five positions during training. A network's capacity for 'spatial generalisation' meant that, at test, the network produced the correct output when object #51 was presented in non-trained positions.

The networks were trained on 50% single stimuli and 50% double stimuli. An epoch of training consisted of one presentation of each acceptable single object pattern, intermixed with an equal number of randomly chosen double object patterns. Weights were adjusted after each training pattern. For consistency, all networks were trained for 5000 epochs.

The dorsal network consisted of the visual input, the dorsal hidden layer, and the dorsal output. The ventral network consisted of the visual input, the two ventral hidden layers, and the ventral output. For the combined network, we connected an intermediate hidden layer laterally with each of the other hidden layers. All networks started with their weights fully randomised.

2.6. *Parameter selection*

We searched for networks that would generalise learned information to novel locations. In this effort, we sometimes varied the number and connectivities of the hidden layers. However, the main parameter we changed was the number of units in each hidden layer.

A network's interpretation of data it has trained on determines how it will act in the face of inputs it has never seen before. A backpropagation-trained network with at least one hidden layer will always try to capture an input–output relationship as well as possible, but the particular way it interprets and represents this relationship can be affected by the number of hidden layers and how big each hidden layer is. Smaller hidden layers force a network to extract more general rules underlying the input–output transformation, while larger hidden layers might allow the network to form a representation for each input it trains on. The 'rules' that a network learns for its input–output transformation guide it when it deals with novel inputs (for related discussion, see Baum and Hausler 1989).

3. Network properties

3.1. *Dorsal network*

3.1.1. *Internal representation and generalisation*

The dorsal network transformed a spatially positioned set of features, as input, into spatially positioned, distributed, feature-based output. Hidden layer sizes which we found to produce good generalisation behaviour were large, in the range of $N = 144$.

Because the input–output transformation was based on systematic rules, we found that if the network learns the input–output transformation for the majority of the objects, it easily generalised that information and responded correctly even to a completely novel object, as well as to object #51 at novel positions.

As the dorsal network generalises across configurations of objects' features for each respective point in space, it would not need to have a spatially invariant representation. More, to prevent interference between objects at different locations, entirely distinct representations for objects at different locations might be adaptive. This is just what we find: The average correlation between hidden layer activations for a given object presented at different positions was 0.064 (SD = 0.106). As a result, the dorsal network was able to process double stimuli with virtually zero error (the sum of squared deviations from the target pattern averaged 0.05 per trial).

In sum, the dorsal pathway learned to generalise across objects, and it seems that this generalisation was learned separately for each position in space. The dorsal network responded correctly to objects in novel positions and even entirely novel objects, and could respond to double stimuli presentations virtually flawlessly.

3.1.2. *Emergent object identity*

The dorsal network did not explicitly represent object identity in its output, and, due to the distinct representations produced in the dorsal hidden layer for an object at different positions, the dorsal hidden layer did not seem to directly represent object identity. However, we wanted to test whether object-identifying information might exist within the dorsal section's internal representation. We froze the weights on a fully trained dorsal network and used its hidden layer activations as input for a network with one hidden layer and an object identity output. This network both correctly recognised the objects it was trained on and could also recognise the 51st object in positions where neither the original part of the network nor the new hidden and output layers had ever seen it during training.

The way in which this could occur is slightly complicated. For each position, the dorsal hidden layer learned about the similarities between different configurations of features. The object identity hidden layer may have learned to extract information about the object features present regardless of their positions in the input layer. Thus, even though the object identity hidden layer had not learned about a particular object for a given position, it could have learned about all of the other objects at that position and so could generalise to the novel item and activate representations for the appropriate features. The weights between the object identity hidden and output layers may have served to transform a particular configuration of features into a given object identity.

These ideas could explain two data points: First, given an entirely novel object, the semantic output could not activate the particular unit that represented the object's identity because all the weights to its untrained output unit were zero. However, it did partially activate the output units for other objects sharing similar features, indicating that it represented the object's configuration of features. In the case of object #51, the network had been trained on object #51 for the first five positions and so the semantic output layer could have learned the configuration of features associated with #51's output unit. When #51 appeared in a different position, it could have reactivated the representation of its configuration of features in the object identity hidden layer and so activated #51's output unit.

The representation in the dorsal hidden layer only required one additional hidden layer to support the identification of objects, in contrast to the two hidden layers needed in the ventral section (described below). We suspect, therefore, that the dorsal hidden layer is not simply reproducing the localist representation of objects in the input layer. Indeed, the relationships between the inputs and dorsal outputs were such that the dorsal section could not select the correct output units based on only an individual feature of an object, but also had to take into account the other features of that object. Achieving the correct outputs likely required that the dorsal hidden layer represent conjunctions of the features active in the input layer.

3.2. *Ventral network*

3.2.1. *Internal representation and generalisation*

In contrast to the dorsal network, the ventral network had to transform information about the features of an object into a localist, spatially invariant representation of the object's identity and category. Positional information was irrelevant to the output of the network. As the outputs were idiosyncratic for each object, the network could not generalise to novel locations by generalising across objects. The solution we found for spatial generalisation was to have two very small hidden layers (e.g. $N_1 = 20$ and $N_2 = 60$), which forced the network to compress its representation by increasingly stripping away positional information.

Compared with the dorsal network, the first hidden layer was much more spatially invariant; the average correlation of hidden layer activations for a given object at different positions was 0.541 (SD = 0.126). This correlation was steady across locations, in that the representation of an object at one location was as similar to its neighbour as it was to a distant position. It is possible that the second hidden layer used this constancy to extract a representation that was even more spatially invariant. By the time the input information was represented in the second hidden layer, the correlation between a given object at different positions was 0.988 (0.004).

A grid search of networks with different hidden layer sizes revealed a tradeoff between the ability to generalise across space and the capacity to process multiple objects: In general, fewer units in the hidden layers led to better spatial generalisation, but smaller hidden layers also led to worse performance at correctly recognising the identities and categories of multiple stimuli. Across 455 simulations, these two characteristics were inversely related, with a Pearson's correlation coefficient of $r = -0.386$ (453 df, $p < 0.001$). Individual network performance varied; further research is necessary to understand this in more detail.

3.2.2. *Categorisation*

With a representation in the second hidden layer that was largely spatially invariant, the ventral network represented idiosyncratic information (such as category) in a way that could generalise across locations. That is, if an object was trained for its identity and category at one location, then when that object was presented at a novel location, its representation might have been funnelled into the same, spatially invariant representation and so activated the same population of units representing its identity and category.

3.3. *Dorsal–Ventral network*

The integrated Dorsal–Ventral network was formed by connecting the hidden layers for the dorsal and ventral sections through an intermediate hidden layer (Figure 1). This D–V network could generalise object identity, category information, and actions outputs across space, and could correctly recognise the identities of double stimuli: Given 1000 double object trials, excluding untrained objects or positions, the correct object identity units were the two most active output units on 96.5% of trials, compared with an average of 26.8% for ventral-only networks with ventral hidden layers of the same size.

4. Illusory conjunctions

Our simulation of illusory conjunctions in normal subjects is based on an experiment by Mozer (1983). Here we examine the propensity of the network to miscombine two stimuli with shortened processing times.

4.1. *Experimental data*

In his first experiment, Mozer (1983) explored the frequency of illusory conjunctions between two similar four-letter words. For a given initial word (the stimulus, e.g. 'LIKE'), Mozer found two complements (e.g. 'LOVE' and 'LANE') that shared the first and last letters with the stimulus and that, miscombined with letters from the stimulus, also formed words (e.g. 'LIKE' and 'LOVE' could be miscombined to form 'LIVE', while 'LIKE' and 'LANE' could form 'LAKE'). Mozer presented the stimulus with either of its complements in two separate blocks of trials.

To control for the effect of errors due to guessing, Mozer calculated the 'true' rate of perceptual crossovers as the proportion of responses combining the letters of the two presented words (the S and Cp) minus the proportion of responses combining the stimulus and the other possible alternatives, but not currently presented, complementary word (Cnp). All other responses were ignored. Mozer set the presentation duration for each subject such that the overall error rate was roughly 30% (for Mozer's results, see Table 1).

With these brief presentations, Mozer found significantly more incorrect responses combining letters from the stimulus and its presented complement (Cp) than errors combining the stimulus and its non-presented complement (Cnp).

4.2. *Simulation results*

Objects representing the stimulus and its complements (Cp and Cnp) were selected from our training set with the criterion that they share one stimulus dimension and differ along both of the

Table 1. Simulation and adapted results from Mozer (1983).

% Responses	Mozer	Simulation
Correct	68.51	74.90
Crossover	13.70	16.88
S-Cnp Error	5.57	0.00
'True' crossover	8.13	16.88

Note: The data in column 1 are from Mozer (1983, p. 534).

others. To simulate the brief exposure durations in Mozer, we presented the D–V network with an object (S) and its complement (Cp), and we found the point in time when activation of the correct object output units surpassed alternative unit activations across 70% of trials.

We found a substantial proportion of trials with erroneous responses that combined the two presented objects (S and Cp), the so-called illusory conjunctions (Table 1). We did not find any responses combining the stimulus and the non-presented complement (S and Cnp). Non-illusory conjunction errors were generally combinations of features from the two presented objects with some non-presented features. One explanation for the different rate of 'guessing' (S-Cnp) errors in Mozer might be mnemonic interference, which our model did not have a mechanism to capture.

5. Dorsal versus ventral lesions

We compare the performance of our model after dorsal and ventral lesions. Lesions of dorsal hidden units strongly affected the accuracy of ventral outputs when two objects were present, suggesting that dorsal representations played a role in ventral processing for multiple object trials. These dorsal lesions led to high rates of illusory conjunctions, similar to the performance of the simultanagnosic patient R.M. (Robertson et al. 1997). Conversely, lesions of ventral layers had minimal impact upon dorsal functioning. This result is similar to the retained abilities of the visual form agnosia patient, D.F., (Milner and Goodale 2006) and semantic dementia patients (Hodges, Bozeat, Lambon Ralph, Patterson and Spatt 2000) to perform basic manipulations of objects despite impaired object recognition.

5.1. *Dorsal pathway damage: R.M. and simulation*

In June 1991, and again in March 1992, R.M. suffered strokes which left him with large bilateral lesions in his parietal cortex and rendered him simultanagnosic. The experiments we report here from Robertson et al. (1997) started roughly 15 months after R.M's second stroke and continued for 15 months, at which point R.M. suffered a third brain injury.

While all of the tests in Robertson et al. involved the simultaneous presentation of two coloured letters, other parameters varied between experiments. We averaged the rates of correct responses and illusory conjunctions for each time point (Figure 3), with two exceptions: two tests involving a white 'cue' to guide responses both had aberrant rates of response. It is possible that this cue, as a third presented object, rendered the target objects more difficult to perceive. As such, the averaged data per session were calculated without the white cue tests.

We simulated this posterior parietal damage by effectively removing different percentages of the dorsal hidden layer units. We did this by creating strong negative biases for a randomly chosen subset of the units, effectively preventing these units from reaching an activation greater than 0. We then froze the negative bias weights and retrained the network with its regular training regimen. We considered the most active ventral output unit at the end of a trial to be the response.

As with R.M., we found an initially large rate of illusory conjunctions (Figure 3). With retraining, we also found gradual improvement in the accuracy of the most strongly perceived object, even for very extensive lesions. These network lesions reveal the ventral reliance upon dorsal processing in the full D–V network when two objects are present.

5.2. *Ventral pathway damage*

Carbon monoxide poisoning in 1988 led D.F. to develop visual form agnosia. Milner and Goodale (2006) assert that this lesion likely interrupted her ventral visual pathway. D.F. was impaired in

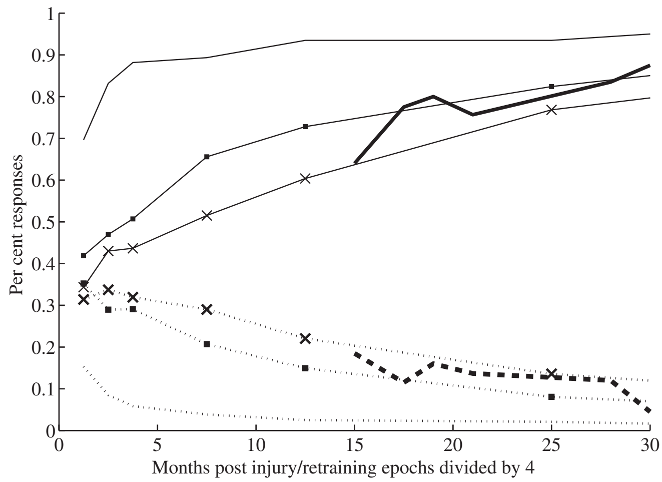

Figure 3. R.M's data (in bold) and simulation results for simultanagnosia. Solid lines are per cent correct responses, dashed lines are per cent illusory conjunction responses. Crosses indicate that 100% of dorsal hidden units were lesioned; squares, 80%; and non-marked, 60%. R.M.'s data extracted from Robertson et al. (1997, p. 304). Copyright 1997 by the Massachusetts Institute of Technology. Adapted with permission.

recognising common objects based on visual shape information alone, although she could identify similar objects tactually. However, D.F. displayed preserved abilities to interact with objects. She was able to rotate objects to fit into a slot whose orientation she could not, separately, describe; and she displayed normal anticipatory finger apertures in grasping different-sized objects that she could not visually discriminate (Milner and Goodale 2006).

Semantic dementia patients demonstrate a related dissociation of skills. Semantic dementia is a degenerative disorder affecting temporal cortex, and leads to progressive impairments in identifying objects. Unlike visual form agnosia, these impairments are not modality-specific (Hodges and Patterson 2007). Potentially related to their object recognition errors, semantic dementia patients also do not tend to grasp common objects in their customary ways. However, semantic dementia patients grasp and interact with novel objects similar to control subjects, suggesting visual information could still guide general motor interactions (Hodges et al. 2000).

To simulate these data, we completely removed the ventral hidden layers from a fully trained D–V network. This resulted in the loss of 87 units from the network; in contrast, the 60% dorsal lesion removed 86 hidden units.

Dorsal outputs were minimally affected, even for double object presentations. The average sum of the squared error was 0.599 per pattern across 1000 double object presentations, where target outputs had an average summed activation of about six. Visual inspection revealed that over 1/3 of double object trials were not substantially affected. On the remaining trials, errors mainly involved one or two units inappropriately activated or inappropriately inactivated. A single epoch of retraining reduced the average sum of squared error to 0.110, and qualitative output errors were very rarely apparent to visual inspection.

6. Conclusion

We have attempted to trace out the emergent properties of a system which produces both systematic actions and object identification. We chose parameters only such that the system would be able to recognise and act upon known objects in novel positions. The localist, idiosyncratic outputs of the ventral section and the distributed, systematic outputs of the dorsal section led our networks to produce two distinct and complementary modes of representation, a more spatially invariant representation of unique object information and a more spatially specific representation of generalities across objects. Trained together, these two systems dynamically cooperated to determine the identities of multiple objects, and could not only replicate the correct perception of two items for humans but also some of the types of errors produced with brief presentation times or neural damage.

We have treated the relationship between visual features and actions as entirely systematic. However, in acting upon an object, visually based knowledge about its shape is often modified by our idiosyncratic knowledge about the object's properties or customary usage. One direction of future work is to test whether our combined network can utilise the ventral section's idiosyncratic knowledge about objects to produce object affordances based on both visual features and semantic information.

There are many further issues to be resolved and further refinements to be made. We ourselves have certain questions that we intend to explore, e.g. how does the dorsal section help the ventral section, and how can we best characterise the behaviour of the middle hidden layer (Figure 1)? How do the network properties change between the independent dorsal and ventral networks and the combined D–V network? What are some possible alternative architectures? We hope to address these and additional questions in future work.

Models of the binding problem often conceptualise the dorsal stream's involvement, if any, either as a source of a synchrony-producing binding signal or as a form of spatial attention that enhances perception for a selected position or positions. For the first, our model did not have a mechanism for oscillations or synchrony and did not produce its results in this way. For the second, in our minds, there is only a fine distinction between saying that one system has an attentional effect on another versus saying that the systems influence each other. Thus our model can be seen as being partially aligned with an attentional account. However, the type of information coming from the dorsal system differed between our model and many models in this area. In our model, 'what' and 'where'/'how' sections complement each other not only in the output they produce but also in their differing internal representations of object information. With rich object information and progressively less spatial resolution across the two hidden layers in one section, and a high degree of spatial information but a less explicit representation of object identity in the other section, it was in the cooperation between these complementary processing systems that our network was capable of correctly perceiving more than one object in parallel.

Acknowledgements

Partial funding for this research was provided by the National Science Foundation (IGERT grant 0801700).

References

Baum, E.B., and Hausler, D. (1989), 'What Size Net Gives Valid Generalization?', *Neural Computation*, 1, 151–160.

Coslett, H.B., and Saffran, E. (1991), 'Simultanagnosia: To See But not Two See', *Brain*, 114, 1523–1545.

Hodges, J.R., Bozeat, S., Lambon Ralph, M.A., Patterson, K., and Spatt, J. (2000), 'The Role of Conceptual Knowledge: Evidence from Semantic Dementia', *Brain*, 123, 1913–1925.

Hodges J.R., and Patterson K. (2007), 'Semantic Dementia: A Unique Clinicopathological Syndrome', *Lancet Neurology*, 6(11), 1004–1014.

Konen, C.S., and Kastner, S. (2008), 'Two Hierarchically Organized Neural Systems for Object Information in Human Visual Cortex', *Nature Neuroscience*, 11(2), 224–231.

Milner, A.D., and Goodale, M.A. (2006), *The Visual Brain in Action* (2nd ed.), New York: Oxford.

Mozer, M.C. (1983), 'Letter Migration in Word Perception', *Journal of Experimental Psychology: Human Perception and Performance*, 9, 531–546.

Remus, D.A. (2010), ‘Learning Position-tolerant Object Representations’, unpublished Ph.D. dissertation, Stanford University, Dept. of Psychology.

Robertson, L., Treisman, A., Friedman-Hill, S., and Grabowecky, M. (1997), ‘The Interaction of Spatial and Object Pathways: Evidence from Balint’s Syndrome’, *Journal of Cognitive Neuroscience*, 9, 295–317.

Treisman, A. (1998), ‘Feature Binding, Attention and Object Perception’, *Philosophical Transactions of the Royal Society B*, 353, 1295–1306.

Treisman, A., and Schmidt, H. (1982), ‘Illusory Conjunctions in the Perception of Objects’, *Cognitive Psychology*, 14, 107–141.