Connection Science, Vol. 14, No. 1, 2002, 1–48

A connectionist approach to processing dimensional interaction

ADRIAAN G. TIJSSELING and MARK A. GLUCK

Cognitive Neuroinformatics Group, National Institute of Advanced Industrial Science and Technology, AIST Tsukuba Central 2, 1-1-1 Umezono, Tsukuba 305-8568, Japan
email: adriaan.tijsseling@aist.go.jp

Abstract. The difference between integral and separable interaction of dimensions is a classic problem in cognitive psychology (Garner 1970, *American Psychologist*, **25**: 350–358, Shepard 1964, *Journal of Mathematical Psychology*, **1**: 54–87) and remains an essential component of most current experimental and theoretical analyses of category learning (e.g. Ashby and Maddox 1994, *Journal of Mathematical Psychology*, **38**: 423–466, Goldstone 1994, *Journal of Experimental Psychology: General*, **123**: 178–200, Kruschke 1993, *Connection Science*, **5**: 3–36, Melara *et al*. 1993, *Journal of Experimental Psychology: Human Perception & Performance*, **19**: 1082–1104, Nosofsky 1992, *Multidimensional Models of Perception and Cognition*, Hillsdale NJ: Lawrence Erlbaum). So far the problem has been addressed through *post hoc* analysis in which empirical evidence of integral and separable processing is used to fit human data, showing how the impact of a pair of dimensions interacting in an integral or a separable manner enters into later learning processes. In this paper, we argue that a mechanistic connectionist explanation for variations in dimensional interactions can provide a new perspective through exploration of how similarities between stimuli are transformed from physical to psychological space when learning to identify, discriminate and categorize them. We substantiate this claim by demonstrating how even a standard backpropagation network combined with a simple image-processing Gabor filter component provides limited but clear potential to process monochromatic stimuli that are composed of integral pairs of dimensions differently from monochromatic stimuli that are composed of separable pairs of dimensions. Interestingly, the responses from Gabor filters are shown already to capture most of the dimensional interaction, which in turn can be operated upon by the neural network during a given learning task. In addition, we introduce a basic attention mechanism to backpropagation that gives it the ability to attend selectively to relevant dimensions and illustrate how this serves the model in solving a filtration versus condensation task (Kruschke 1993, *Connection Science*, **5**: 3–36). The model may serve as a starting point in characterizing the general properties of the human perceptual system that causes some pairs of physical dimensions to be treated as integrally interacting and other pairs as separable. An improved understanding of these properties will aid studies in perceptual and category learning, selective attention effects and influences of higher cognitive processes on initial perceptual representations.

Keywords: integral and separable dimensions, connectionism, Gabor filters.

Connection Science ISSN 0954-0091 print/ISSN 1360-0494 online © 2002 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/0954009021013859 4

1. General introduction

One of the basic tasks humans have to do continuously is to sort things, objects, or events into distinctive groups based on the similarities between the various instances. This process of categorization involves finding or constructing the feature(s) that distinguish the members of each category. Since these features may vary in several dimensions, a successful theory of categorization needs to explain how these different dimensions interact. Previous studies suggest that there are at least two different types of interaction (Garner 1970). The interaction of a pair of dimensions can be integral, meaning that they are perceived holistically. Goldstone (1999, 2002) refers to the underlying dimensions as being psychologically fused. A standard example is the brightness and saturation of a colour. When one of these features changes, the effect is perceived as a change in overall colour, rather than as a change in one attribute of that colour. More commonly, a pair of dimensions is perceived analytically as separable. For example, the size of a square and the brightness of its colour are separable dimensions: if the brightness is changed, we do not normally perceive a difference in size.

The problem of integrality versus separability has been an important aspect of theories of categorization. Researchers have been approaching the problem by testing human categorization performance (e.g. Garner and Felfoldy 1970), devising statistical models to fit experimental data (e.g. Nosofsky 1986, Kruschke 1993), or applying direct similarity scaling (e.g. Carroll and Arabie 1980). A typical experiment consists of presenting stimuli that are composed of a pair of dimensions, such as brightness and saturation of a colour, to human subjects with the instruction to sort them, identify them, or discriminate between them. The performance of the subjects is analysed, usually by measuring the amount of time it takes them to sort a set of stimuli or by observing the errors in discrimination and identification. A general finding is that performance significantly differs between integral and separable dimensional interaction (e.g. Garner and Felfoldy 1970).

An example of a statistical model is the Generalized Context Model (GCM), proposed by Nosofsky (1986), which uses psychological similarity ratings as input. According to the GCM, a stimulus belongs to a category if the summed similarity of this stimulus to all stored category exemplars exceeds a certain probability estimate. Stimuli are represented as points in a multidimensional space, with the similarity between any two stimulus representations described as a decreasing function of their distance in that space. These interstimulus distances are derived from multidimensional scaling results from human subjects' similarity ratings of pairs of actual stimuli (see also Shepard 1987). The difference between integral and separable interaction is then incorporated into how distances between stimulus representations are calculated.

Although the above approaches to the problem of dimensional interaction have produced relevant data, they remain descriptive and *post hoc*. Models such as the GCM do not explain how stimulus representations are formed. They are used to fit behavioural categorization data based on acquired human similarity ratings. The question remains: What mechanism underlies the transformation from a raw stimulus to an internal representation and how does this transformation affect the interaction between dimensions? After several decennia of analysis it is still not clear how and why integral pairs of dimensions are processed differently from separable pairs. It might be the case that the behavioural data are simply insufficient to constrain or guide appropriate theories of dimensional interaction without the benefit of computational modelling.

A different approach might be needed to find an explanation for the differential processing of integral and separable dimensions, one that provides insights into the mechanism involved in the differential impact of separable and integral interaction of dimensions on category learning. Within this approach, the focus would be on exploring how distances between stimuli are transformed from physical to psychological space when learning to identify, discriminate and categorize them. In this paper, we provide theoretical arguments for our hypothesis that connectionist models are a possible candidate for a mechanism of differential processing of dimensions. Although connectionist models are not yet as powerful as human categorizers, we shall show that combining a neural network with a suitable sensory-preprocessing input layer will create a basic capacity to process differentially stimulus dimensions. An analysis of the way these models categorize can in turn illuminate the way in which it might be done in the human cognitive system and provide a new perspective on human subject data. An advantage of a mechanistic approach to dimensional interaction is that it takes the same spatially organized stimuli that are presented to subjects rather than *ad hoc* inputs.

The idea of attaching a sensory preprocessing component in a connectionist model finds support in a movement within the cognitive sciences that emphasizes the important bidirectional relation between perception and cognition. Goldstone (1998a) argues that perception contains an initial source of structured information that can be operated upon subsequently by higher cognitive processes (Tijsseling 1998), yet these cognitive processes can in some cases also modify percepts (Goldstone 1995, Goldstone *et al*. 1997, Schyns *et al*. 1998). In particular, perceptual learning and categorization are constrained by the existing structure of the sensorimotor apparatus of an organism. These constraints allow for adaptation to the environment and serve as a starting point for the development of more sophisticated percepts, because they determine what can and cannot be perceived and, consequently, learned (Grossberg 1982, 1987, Karmiloff-Smith 1992, Murre *et al*. 1992, Goldstone 1998a).

Given this bootstrapping of perception by the sensorimotor system, we shall argue that if the encoding of an input to a neural network is motivated by the way human-like perception encodes the physical structure of the stimuli, then the question whether the dimensions that compose the stimulus in question are interacting in an integral or separable manner can be determined without any *ad hoc* or *post hoc* fitting by an external observer. The interaction between a pair of dimensions that make up a stimulus is not determined by the physical structure of the stimulus in question, but is based on how the human perceptual system processes the physical stimuli. A Martian might have a different kind of perceptual apparatus and perceive the same stimulus in a radically different way. In other words, the physical structure of the stimulus is the same, but the way it is reconstructed within a cognitive system, or which aspects of information are extracted from it by the system, might differ across species.

The potential of connectionist models for explaining the differential processing of integral and separable dimensions is demonstrated in this paper by simulations with a backpropagation network combined with an image-filtering input layer composed of Gabor filters (Gabor 1946, Daugman 1988). Backpropagation networks have been used extensively in categorization models (e.g. Kruschke 1993, Gluck and Myers 1993, Cangelosi *et al*. 2000). We are using backpropagation only as an illustration of our theoretical arguments, showing how a simple neural network may already employ a crude mechanism for differential processing of dimensional interaction. As a

candidate for raw stimulus processing, we have opted for a Gabor filter model of sensory processing in the primary visual cortex (Marcelja 1980). By using both backpropagation and Gabor filters (hereafter simply referred to as the model), we try to combine a mechanism for transforming a raw stimulus into a psychological representation together with a feature filtering mechanism for category learning. For the purpose of explanation, we have tried to keep the model intelligible while still going a long way toward explaining the basis of dimensional interaction.

We shall begin by reviewing empirical evidence for a distinction between integral and separable pairs of dimensions and then we shall discuss how these two kinds of interaction can be interpreted in the context of physical versus psychological space. For this we refer to a seminal paper by Shepard (1964) that deals with the transformation of distances between stimuli from physical to psychological space and how multiple dimensions interact in similarity judgements. We shall describe our demonstrative model and apply it to experimental studies concerning differential processing of dimensions. It will be shown that the model's performance is qualitatively similar to human subjects. We shall explain how the model addresses the problem of dimensional interaction and argue that this may suggest a general underlying mechanism in which distances between representations of stimuli in psychological space are derived from the corresponding physical distances. In particular, we shall single out the crucial role Gabor filter encoding plays in the processing of dimensional interaction. Finally, we shall discuss the lack of an attention mechanism in the model, which prevents it from explicitly attending to one single separable dimension. We offer a solution to this problem by injecting a crude selective attention mechanism to the model, based on previous work by Kruschke (1996). With an attention mechanism in place, the model processes the paradigm task of condensation versus filtration in a manner qualitatively similar to human subjects.

2. Integral versus separable dimensions and their relation to isosimilarity contours

In typical categorization studies, stimuli are often employed that vary in several continuous dimensions. To complicate matters, a subject's perception of how dimensions interact may vary for each different pair of dimensions. At the one extreme, we have integral pairs of dimensions (Garner 1970), such as, for example, brightness and saturation of a colour, which tend to be holistically perceived. Subjects who have to evaluate the brightness of a stimulus suffer interference (i.e. speed and accuracy deficiencies) if saturation is varied at the same time (Torgerson 1958, Garner and Felfoldy 1970, Handel and Imai 1972). At the other extreme, there are separable pairs of dimensions (Garner 1970), such as brightness and size: subjects who have to focus on one of these two dimensions can do so even when the other irrelevant dimension is varied (Attneave 1963, Handel and Imai 1972, Garner 1974, Gottwald and Garner 1975). In short, subjects can attend to each of the separable pair of dimensions separately, while integral pairs of dimensions appear to be perceived as if they are 'psychologically fused'[1] (Goldstone 2002). Integral and separable pairs of dimensions have been found in the visual, auditory and vibrotactile modalities (Garner 1970). Hence, the distinction between integral and separable pairs of dimensions is fundamental for models of human categorization and continues to generate active research interest (e.g. Melara *et al.* 1993, Ashby and Maddox 1994).

More accurate judgements of the interaction between dimensions can be made

using psychological distances. Several studies have shown that there is a relationship between the type of interaction between dimensions and the metric that fits these psychological distances. Shepard (1964) focused on the question of how differences in two dimensions combine psychologically into an assessment of the overall similarity between two stimuli. This cannot be answered by just looking at the relation between the physical stimuli themselves, because psychological similarity is not just dependent on physical similarity. For example, in the case of colour, the similarity between one green colour and another and between the same green colour and a blue one is very different psychologically, even though the physical distance between the wavelengths might be identical (Berlin and Kay 1969, Bornstein 1987). The only property that seems to be invariant is that when two stimuli approach each other in physical space, then their psychological representations will be more similar (Shepard 1964). In this respect, the issue of assessing overall similarity is to find the transformation that will convert physical interstimulus distances into psychological distances between the corresponding representations.

Discovering this transformation is complicated when more than one dimension is involved, because one can construct a separate psychological scale for each of the underlying physical dimensions, but this will not capture the similarity between two stimuli that vary on several dimensions. In a one-dimensional case, the similarity between any two points on the dimension is symmetrical, but this symmetry is not present when stimuli vary in more dimensions. Consider, for example, figure 1, which shows a point S in a plane spanned by two dimensions *A* and *B*. Given S, we can create a series of stimuli on some imaginary straight line that passes S at a specified distance in physical space (for example, stimuli D, H and L as shown on the left-hand side of the figure). We assume that points D and L have the same amount of physical and psychological similarity to S, but each of which varies from S in just one dimension at some distance, respectively, d(S,D) and d(S,L). Point H, on the other hand, also varies from S, but along both dimensions and at a given distance of $k \cdot d(\mathrm{S,D}) + k \cdot d(\mathrm{S,L})$. We cannot tell how large this factor k must be in order for the psychological similarity between S′ and H′ to be the same as the similarity between S′ and either D′ or L′, this remains a function of how the two dimensions *A* and *B* interact.

Shepard (1964: 56) argues that knowing the rules that produce an overall similarity between stimuli that vary in more than one dimension is equivalent to finding the characteristics of the corresponding *isosimilarity contour*, which is defined as 'the shape of the locus of all stimuli that have any prescribed degree of similarity to any prescribed standard stimulus'. For example, what are all the stimuli that have the same fixed similarity to S in figure 1? In other words, if we know all the stimuli with the same given amount of similarity to S, then what shape or locus would these stimuli form? The form of this shape is strongly related to the kind of interaction between the two dimensions that make up the physical space.

According to Shepard (1964), determining the shape of the isosimilarity contour is equivalent to determining the form of that particular Minkowski metric that seems to be appropriate for the psychological space. Minkowski metrics are defined as:

$$d(i,j) = \left(\sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right|^{r} \right)^{1/r} \quad , r \geq 1$$

in which $d(i,j)$ is the distance between stimuli i and j, x_{ik} is the location of the stimulus

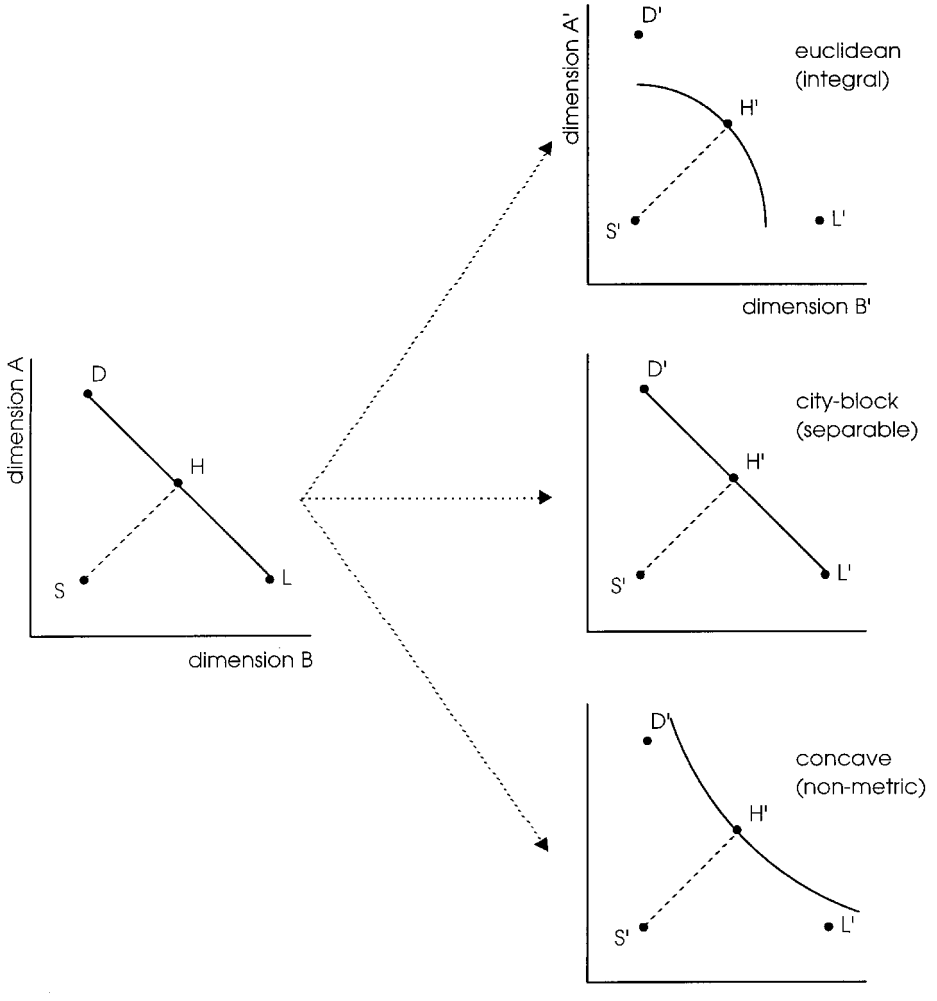


Figure 1. Possible transformations from physical space to psychological space. The figure on the left-hand side displays physical stimuli that vary in two dimensions *A* and *B*. S is a prescribed stimulus and D, H and L are stimuli that lie on an imaginary straight line passing S at a given distance. If stimuli D and L have the same similarity to S, it does not provide any information on the similarity of H. Based on how the dimensions *A* and *B* interact, three resulting isosimilarity contours are shown on the right-hand side: Euclidean, city-block, or concave. S′, D′, H′ and L′ are the psychological space representations of the corresponding stimuli.

i on the *k*th dimension, *n* is the number of dimensions and *r* is the Minkowski parameter. A value of 1 for *r* corresponds to the city-block metric, and a value of 2 denotes the Euclidean metric. Metrics also obey the following principles: the triangle inequality ($d_{ij} \leq d_{ih} + d_{hj}$ for those stimuli *i*, *j* and *h*, where *h* lies between *i* and *j* on a shortest connecting path), positivity ($d_{ij} > d_{ii} = 0$ for $i \neq j$) and symmetry ($d_{ij} = d_{ji}$). Based on these metrics one can produce several possible isosimilarity shapes.

Figure 1 shows that the curve consisting of all stimuli with a specific amount of similarity to S can have different shapes depending on the nature of interaction between the constituent dimensions, *A* and *B*. All of these curves form continuous,

centrally symmetric closed curves as they rotate around S. Based on subjects' judgements of the similarities of these stimuli to S, one can construct the isosimilarity contour for the corresponding psychological space and, consequently, infer the nature of interaction between the dimensions *A* and *B*.

If subjects judge H to be more similar to S than D and L are (figure 1, top), then the isosimilarity contour is elliptical with the latter two stimuli falling outside the contour for H. In this case, the psychological space conforms to a Euclidean metric (Hyman and Well 1967) and the dimensions *A* and *B* can be considered to interact in an integral manner (Lockhead 1972, Garner 1974). Stimuli composed of an integral pair of dimensions are initially perceived holistically, which means that the individual dimensions constituting the stimuli are in effect not perceived. Note that in the case that *r* is equal to 2, rotation of the dimensional axes *A* and *B*, as a consequence, does not change the psychological space.

On the other hand, if subjects judge that D, H, L are equally similar to S (figure 1, middle), then the isosimilarity contour is a straight line and we can conclude that the psychological space obeys a city-block metric, which in turn implies that the pair of dimensions *A* and *B* is separable. In this case, the dimensional axes cannot be rotated, because each dimension can be attended to separately and a rotation would significantly disturb the similarity space of the stimuli.

As mentioned earlier, there are also cases in which the pattern of interaction does not seem to match either integrality or separability, but rather lie in between these two endpoints (Pomerantz and Garner 1973). For these interactions, the appropriate Minkowski metric would be defined by an *r* between 1 and 2. (For a discussion where *r* approaches infinity, see Johannesson (2001).) It is also possible that subjects judge H to be less similar to S than D and L are (figure 1, bottom). This indicates that the isosimilarity contour is concave, which in turn informs us that there is no metric representation of the similarity space, because concave contours violate the triangle inequality rule for metrics.

Given the different metrics, depending on the perceptual distinctiveness of the dimensions of the stimulus, Shepard (1964: 59) argues the necessity of a 'more thorough investigation into the relations of similarity among a set of stimuli that differ in a small number of perceptually distinct and salient dimensions'. In the next section, we shall describe the stimuli used in Shepard's experiment (1964) and a corresponding experiment. Based on this, a series of simulations will be described. The main reason for using data from Shepard (1964) is that Shepard used a well-defined set of stimuli organized in a well-defined two-dimensional stimulus space. In this set of stimuli, adjacent pairs are physically equally distant from each other, so we can make valid inferences about possible isosimilarity contours. In addition, the analyses Shepard provided for the human subject data are extensive and thorough, which makes it much easier to compare with and relate to simulation results. To our knowledge there are no new data superseding or refuting Shepard's seminal work.

3. Description of an experiment from Shepard (1964)

Shepard's experiment measured the number of confusions made by subjects when learning to respond with a unique identifying label for each stimulus object. This method would possibly resolve the question of whether the isosimilarity contour of the psychological space is concave and, as a consequence, whether there is a metric representation of psychological space. The stimuli used by Shepard (1964) in his

experiments were all circles containing a radial line. Series of stimuli could vary in the size of the circle, the angle (tilt) of the radial line, or both. The experiment described below is the final experiment described in Shepard's paper; all details regarding experimental set-up and procedure can be found in Shepard (1964).

Eight stimuli were constructed with reference to a prescribed stimulus, S, which is a circle with a diameter of 1.905 cm containing a radial line at an angle of 45.0°. The location of these stimuli in physical space formed the corners of an imaginary octagon, as shown in figure 2. The horizontal axis corresponds to the diameter of the circle (size) and the vertical axis corresponds to the tilt of the radial line. Shepard designed the training stimuli in such a way that each adjacent pair of stimuli varied in either one or both dimensions. For example, stimuli 8 and 1 varied in the dimension of diameter (2.299 cm versus 2.875 cm), but had the same angle (80°). On the other hand, stimuli 1 and 2 varied in both dimensions of size and tilt. This alternating pattern of one-dimensional variance and two-dimensional variance in an octagonal configuration has been shown to be effective for discriminating between elliptical and four-cornered isosimilarity contours, and therefore reliably indicates whether the pair of dimensions that define the stimuli is either integral or separable (Shepard 1964).

The task of subjects in this experiment was to identify each of the eight stimuli, using a paired-associate method of correction, with a unique letter, but they never saw the reference stimulus S itself. The errors subjects made during identification were used to calculate a confusion score. We focus on adjacent pairwise confusions (i.e. the number of times subjects identified a stimulus as its neighbour, for example, when 1 is identified as 2 in figure 2) as these were shown to be the significant data (Shepard 1964). It can be hypothesized that the amount of confusion for stimuli varying in both dimensions should be significantly different from the amount of confusion for stimuli varying in one dimension only, as previously illustrated with figure 1.

In figure 3, the graph of the errors made during identification learning ('confusions'), averaged over 200 stimulus presentations and 20 subjects, is recreated from Shepard (1964). The points in the graph correspond to the average number of confusions for the eight pairs of adjacent stimuli, with the mean number of confusions marked by the horizontal dashed line. The curve shows that subjects tend to confuse adjacent pairs of stimuli that differ in one dimension more often than those stimuli that vary in both dimensions. In particular, this curve shows an alternating repetition of high confusion for stimuli varying in one dimension and low confusion for stimuli varying in both dimensions, an alternation that seems to go through four complete cycles (of one □ and one ◇). This is a strong indication that the dimensions are interacting in a separable manner because, as we argued, an integral pair of dimensions would be processed holistically and therefore would not produce this pattern of results. The alternating character of single dimension variation versus two-dimension variation moreover suggests that the psychological representation of similarities to S does not match an elliptical isosimilarity contour because the two dimensions that made up the stimuli appear to have an additive character: stimuli that varied in both dimensions proved to be easier to identify, since subjects were able to use information from both single dimensions.

In addition, figure 3 shows another remarkable feature. With each pair of adjacent positions around the octagon (starting at the 2–3 pairwise similarity comparison) the number of confusions seems to increase for stimuli that vary in a single dimension and decrease for stimuli that vary in both dimensions, in a linear fashion. This observation is not mentioned in Shepard (1964), but we feel that this may reveal additional

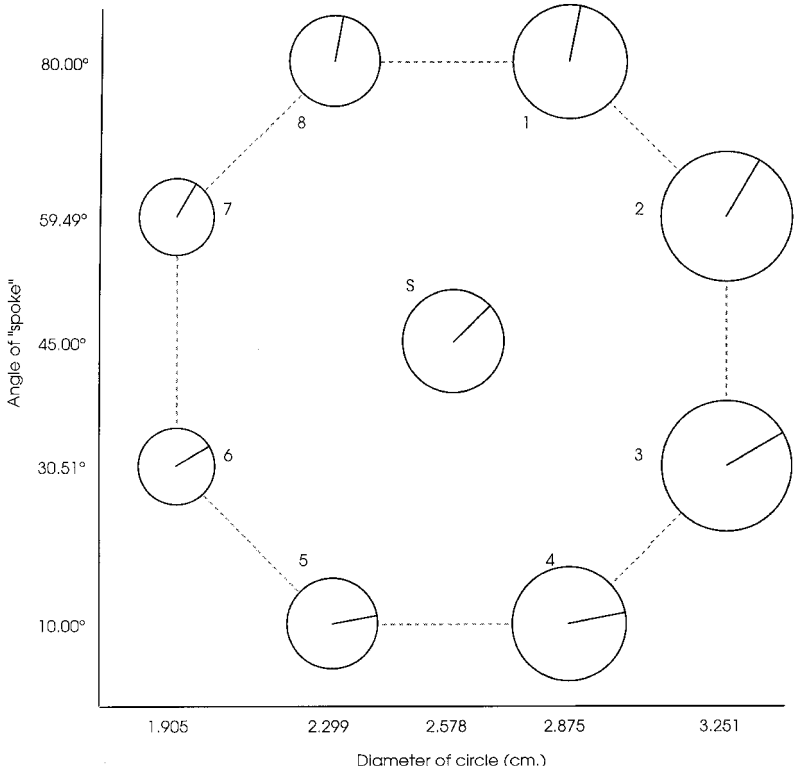


Figure 2. The stimuli in the second experiment were constructed in such a way that their physical similarity to the given standard stimulus S forms an octagon of which S is the centre. The horizontal axis displays the diameter of the circle and the vertical axis displays the tilt of its radial line. (After Shepard 1964.)

properties of the underlying psychological space. Since subjects did not have to rate the similarity between the stimuli and the standard stimulus S, there is no possibility of determining psychological distances to S from these experimental data. Using neural network simulations, however, we might be able to combine evidence to infer the nature of the psychological space and suggest explanations for the observed linear change in error, possibly by exploring the warping of internal representations during learning (Tijsseling and Harnad 1997). In the next section we shall describe the demonstration model and the simulations.

4. Description of the demonstrative model

The model consists of a neural network combined with a Gabor filter input layer (figure 4). We used standard backpropagation with units in the range [0.0, 1.0], sigmoid activation rules and a mean-squared error rule. The network is set up to accommodate Shepard's experiment in which subjects had to identify stimuli by a unique letter. The network had to produce an identifying response as well, i.e. there was a unique output unit, which was to become active for each individual stimulus object. For example, stimulus 1 would produce an output vector of <1,0,0,0,0,0,0,0> and stimulus 2 would produce <0,1,0,0,0,0,0,0>, etc. We shall refer to this procedure as 'identification' and

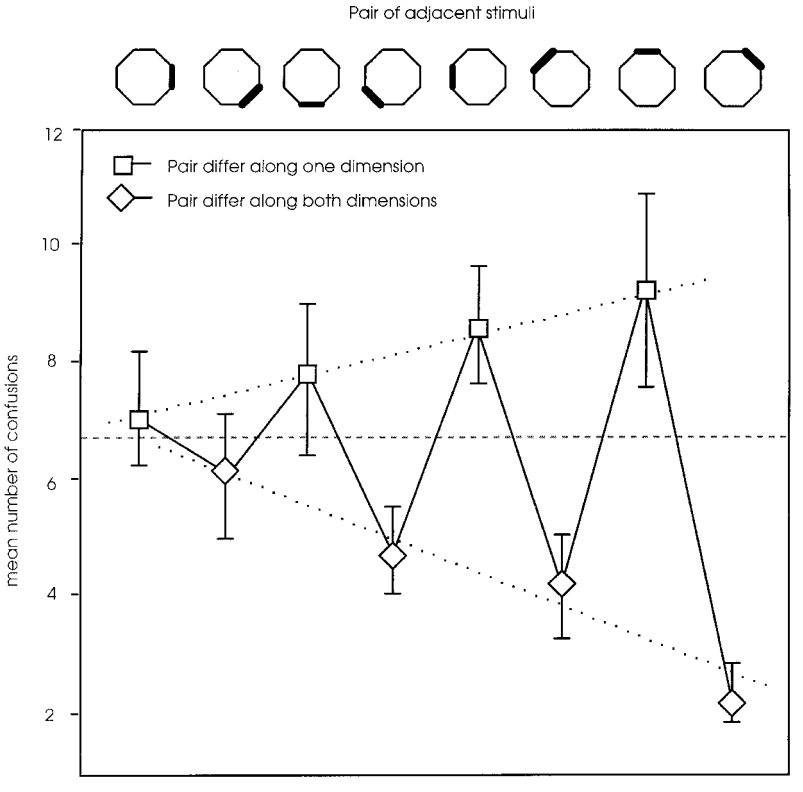


Figure 3. Frequencies with which pairs of adjacent stimuli were confused after learning. The icons on top of the figure correspond to the octagonal configuration of the stimuli as shown in figure 1. The bold line in each icon indicate which pair of adjacent stimuli from the octagonal configuration corresponds to the amount of confusion shown. In addition, confusion scores from stimuli varying in one dimension are marked with a square, whereas confusions from stimuli varying in both dimensions are marked with a diamond. Finally, the error bars indicate the variance over the number of subjects. (Figure recreated from Shepard 1964.)

we intend to embody two crucial processes of category learning using this method (Gluck and Myers 1993, Myers and Gluck 1994): first, stimuli that lead to different responses are separated through predictive differentiation; and second, stimulus features that tend to co-occur are grouped through redundancy compression.

Images of experimental stimuli are processed by several Gabor filters, each tuned to a specific orientation (Gabor 1946, Daugman 1988). Gabor filters have been used successfully for simulations of human face recognition (Padgett and Cottrell 1998, Kalocsai *et al.* 1998), other pattern recognition behaviours (Buse *et al.* 1996) and also categorical perception (Goldstone *et al.* 1996). They are a simplified approximation of human visual processing of non-moving monochromatic sensory stimuli. These filters are similar in shape to the receptive fields of simple cells in the primary visual cortex (V1), which are restricted to small regions of space and are highly structured (Marcelja 1980). Several researchers have described these cells as edge detectors (Hubel and Wiesel 1962, 1965), but their responses can be more accurately described as local measurements of frequencies (Jones and Palmer 1987). Appendix A provides a mathematical description of a Gabor filter and how it has been used in the model to preprocess a monochromatic image. The responses of one or more Gabor filters

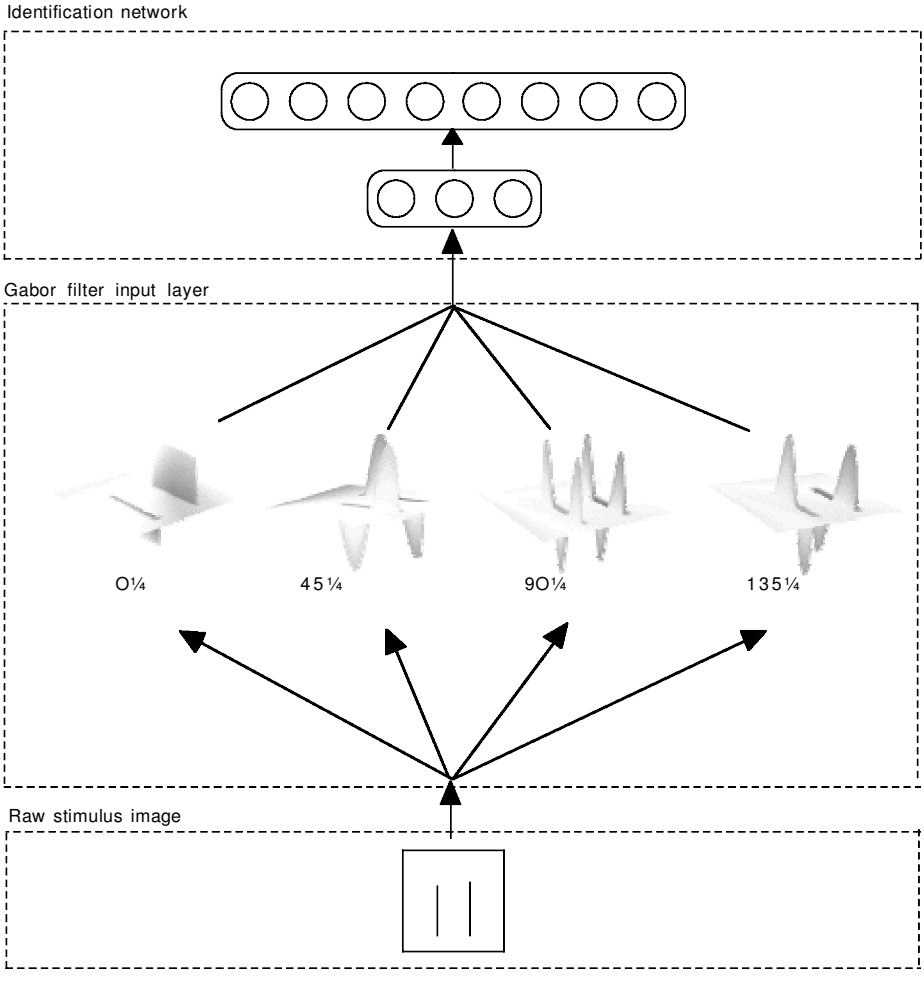


Figure 4. The simple model of dimensional processing and identification and category learning used in this paper. A raw stimulus image is first presented to a Gabor filter layer. This consists of a set of Gabor filters that are tuned to specific orientations. The responses from this Gabor filter layer are presented as inputs to a backpropagation identification network for redundancy compression and predictive differentiation.

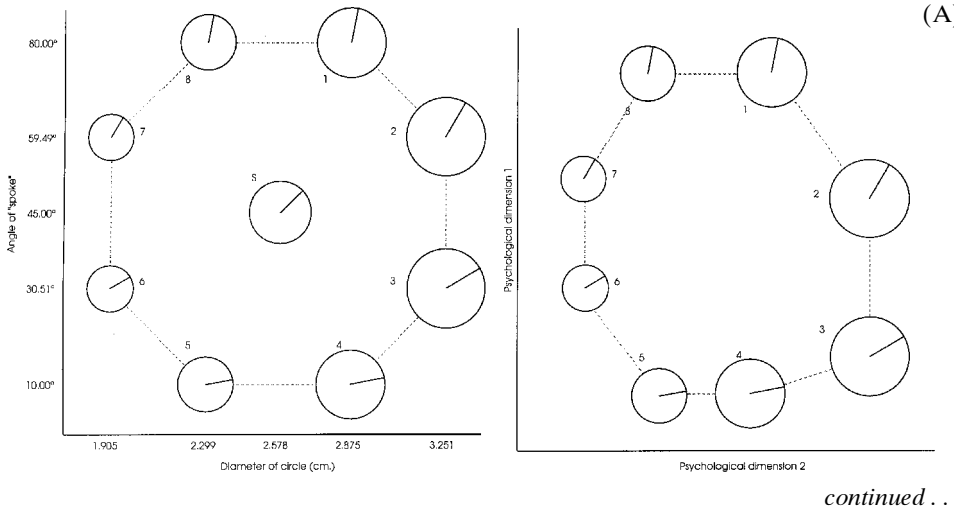
centred at the full image measure frequencies at various locations of the image. Gabor filters, therefore, reflect the physical properties of the stimulus as well as the similarity gradient between stimuli (Goldstone *et al.* 1996, Tijsseling and Harnad 1997). The encoding process constructs a representation of the stimulus by reproducing the intrinsic structure of the stimulus space in a rough approximation of the early stages of human vision.

The responses from the Gabor filters are propagated to the backpropagation identification network, which has to map the sensory representation to a localist encoding of its symbolic label. Stimuli were presented in permuted order. The architecture of the network conformed to an I–H–O hierarchical feedforward structure, in which the size of the input layer, I, was equal to the length of the normalized responses of the Gabor filters used to process the image (ranging from eight to 72, see Appendix A),

the size of the output layer, O, was equal to the number of stimuli, and the size of the single hidden layer, H, was initially set at three, unless mentioned otherwise. The performance for more hidden units was not statistically different. Laakso and Cottrell (1998) show that networks with different numbers of hidden units all achieve essentially identical representational structures given a common stimulus domain. In addition, using three units greatly facilitates analysis of the hidden unit space, which benefits the purpose of this demonstrative connectionist model. In most connectionist approaches, the hidden representation is seen as corresponding to psychological space. In our case, we also used hidden unit activations for analyses.

5. Simulations of the experiment from Shepard (1964)

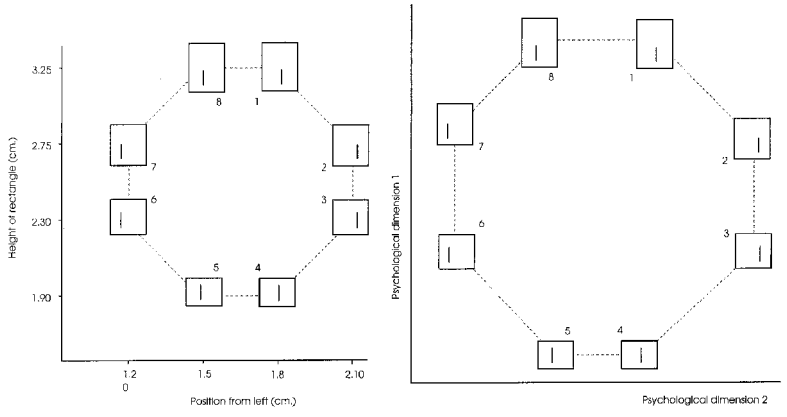
In order to find out whether the results of the combined model are specific to the stimuli employed by Shepard or whether they exhibit some general property of our model, we also trained it on three other sets of stimuli, shown in figure 5. These stimuli were composed of: a separable pair of dimensions of height of a rectangle and position of an embedded vertical line (Kruschke 1993, figure 5(B)); an integral pair of dimensions of height and width of a rectangle (Monahan and Lockhead 1977, figure 5(C)); and an integral pair of dimensions of the length of two parallel vertical lines (Monahan and Lockhead 1977, figure 5(D)). The physical specifications of the stimuli are based on those described in the corresponding papers, but we picked the stimuli in such a way that they formed an octagonal configuration, similar to the configuration of the stimuli of figure 1. The stimuli from Kruschke (1993) were already in octagonal



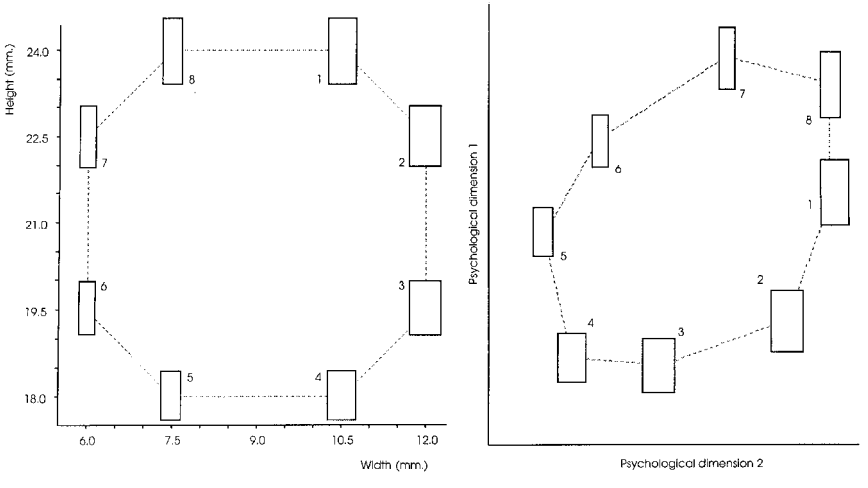
continued . . .

Figure 5. On the left-hand side the four stimulus sets are shown, constructed in such a way that their physical similarity to a given standard stimulus S forms an octagon. (A) The stimuli used by Shepard. (B) Stimuli composed of the separable pair of dimensions of height of a rectangle and the position of an embedded vertical line (Kruschke 1993). The stimuli in (C) and (D) are derived from Monahan and Lockhead (1977) and have integral interacting dimensions: width and height of a rectangle and the length of two parallel vertical lines, respectively. On the right-hand side, the organization of the eight stimuli in psychological space of the model is shown (explained in the text). The dimensions of (A) and (B) interact separably, those of (C) and (D) integrally. The network graphs are obtained from distances between representations in activation space from simulations with two hidden unit networks.

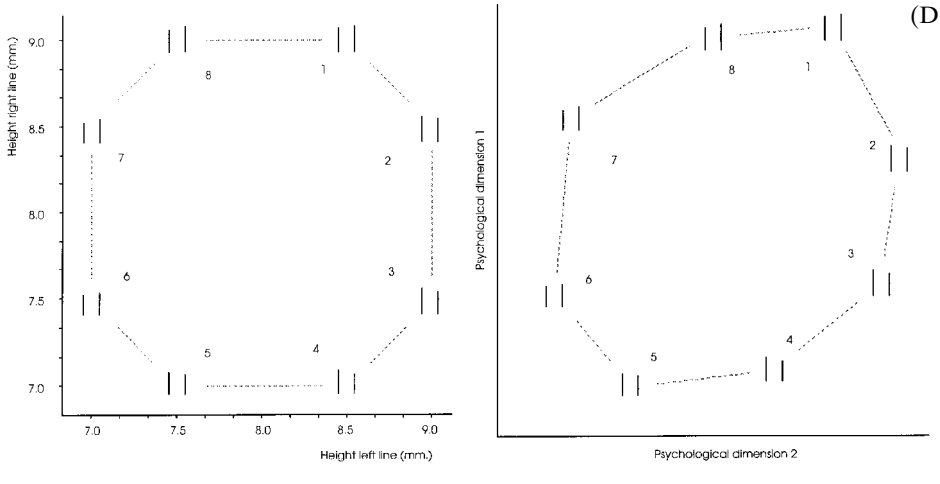
(B)


Height of rectangle (cm.)
3.25
2.75
2.30
1.90
1.2
0
1.5
1.8
2.10
Position from left (cm.)
Psychological dimension 1
Psychological dimension 2

(C)


Height (mm.)
24.0
22.5
21.0
19.5
18.0
6.0
7.5
9.0
10.5
12.0
Width (mm.)
Psychological dimension 1
Psychological dimension 2

(D)


Height right line (mm.)
9.0
8.5
8.0
7.5
7.0
7.0
7.5
8.0
8.5
9.0
Height left line (mm.)
Psychological dimension 1
Psychological dimension 2

configuration and have been kept exactly the same for comparison purposes and also with reference to the filtration/condensation task described in a later section.

These various sets of stimuli will provide us with a view of the kinds of representations—and corresponding psychological spaces—that the network can form. The model is hypothesized to process differentially the various kinds of interaction between dimensions. Hence, the stimuli composed of a separable pair of dimensions as determined by human performance should lead to a similar distribution of interstimulus distances as for the Shepard stimuli, but the stimuli with integral pairs of dimensions should result in a different kind of distribution.

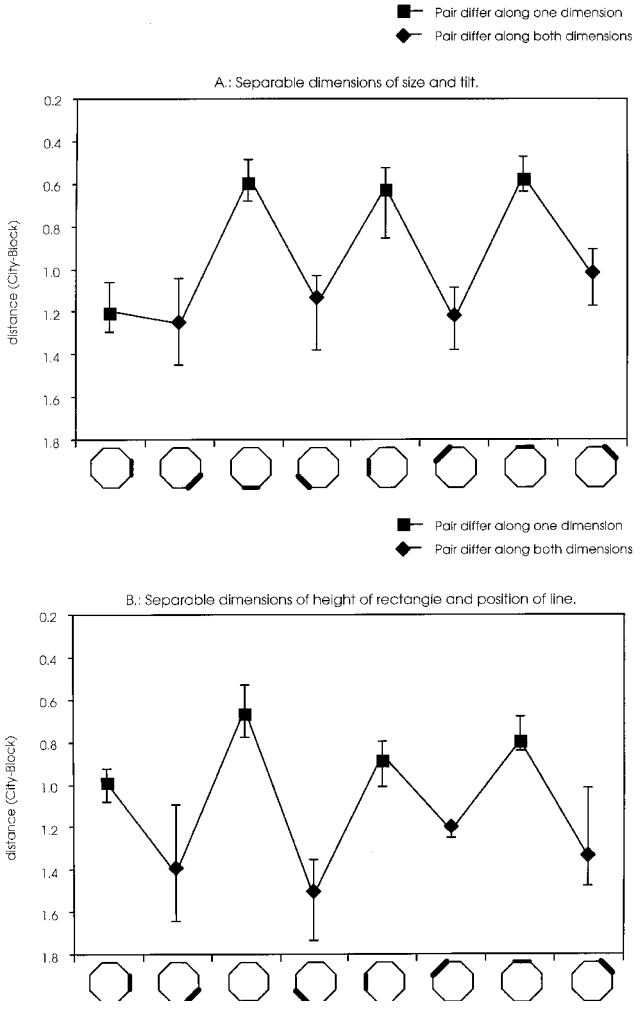
The raw stimulus image is first processed by the Gabor filter component of the model and the normalized responses from the Gabor filters are presented to the backpropagation network (Appendix A). Five repetitions were done, each with a different set of uniformly distributed random initial weights from the range [0.0, 1.0]. The learning rate, β, of the backpropagation network was set to 0.7 and the momentum, α, was held at 0.0. We chose these values as they force the network to learn at a relatively slow rate, which helps it to find the best solution to the task. Other combination of learning rate values (we did tests with values from 0.1 to 0.8 for the learning rate and 0.0 to 0.4 for the momentum in steps of 0.1) did not significantly change the results, but did lead to two networks that had a higher mean-squared error (*MSE*), because they ended up in a local minimum. On average, a low *MSE* was reached after about 1000 epochs (one epoch is one presentation of the entire stimulus set), but to get an accurate view of the organization of representations in hidden unit space we overtrained the network until 10 000 epochs were completed, after which the *MSE* was below a criterion of 0.01. The number of epochs used is not psychologically plausible, but our purpose with this model is qualitative analysis, not quantitative analysis.

We ran one additional simulation to show the significance of applying a Gabor filter encoding that preserves the physical properties of stimuli and mimics human perceptual processes. In this simulation, stimuli were *ad hoc* encoded by using values that describe the magnitude of size and tilt of the circle (for example, 80° would be 0.8 and size 3.251 cm simply converts to 0.03251) because one might argue that it is also possible to just represent the stimuli by their absolute size and tilt values. However, we shall show that the transformation of physical to psychological space is a relevant factor in the differential processing of dimensional interaction. The model has to remain close to what subjects perceive and, just as humans do, it has to extract actively the features from a visual representation of the stimuli, preserving the similarity gradient over stimuli. Dimensional interaction might be embedded in how physical stimuli are processed by a perceptual system, which means that *ad hoc* encoding the stimulus structure to a two-element input vector with values for size and tilt might very well change the nature of the underlying dimensions and trivialize simulation results because they bypass the effects the processing by the perceptual system has on the psychological space. In fact, we shall show that the Gabor filter component has already captured most of the dimensional interaction.

Finally, we assume that there is a direct negative correlation between the identification error (number of confusions) exhibited by the human subjects and the distance between the representations in hidden unit space. That is, if the distance between two representations increases, then the probability they will be confused with each other decreases. Therefore, we calculated the distances between representations in the network's hidden unit activation space as a measure of comparison with the average

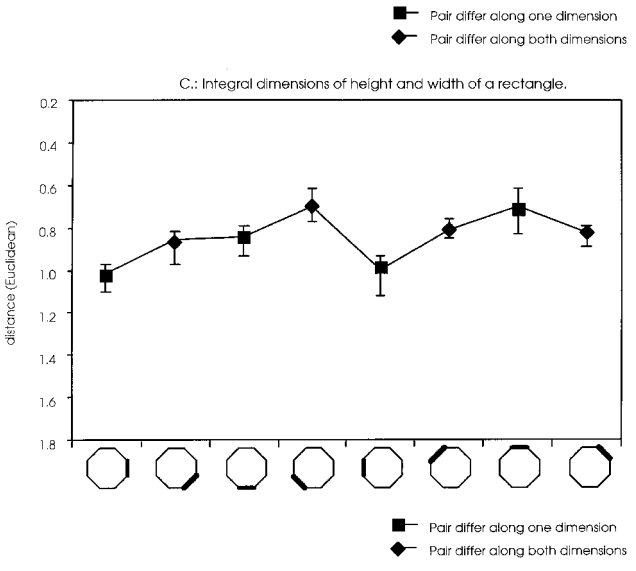
mean number of confusions from Shepard's data.[2] The results of the simulations using the separable stimuli, the integral stimuli and the direct encoding are shown in figure 6, in which error bars denote the minimum and maximum distances between representations in hidden unit activation space over the five networks trained on each stimulus set.

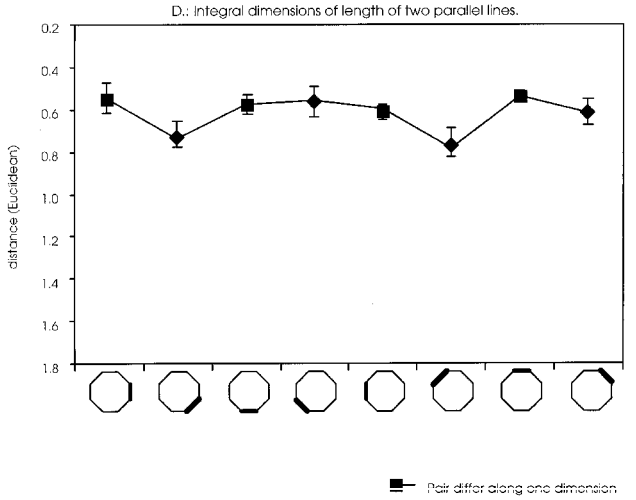
Both graphs for the separable pairs of dimensions show a high confusion score for stimuli that vary in one dimension combined with a low confusion score for stimuli varying in both dimensions. Comparing the pairwise distances between the representations of the network with a set of carefully recovered values from the confusion data of Shepard's experiment for pairs of adjacent stimuli produces a significant Pearson

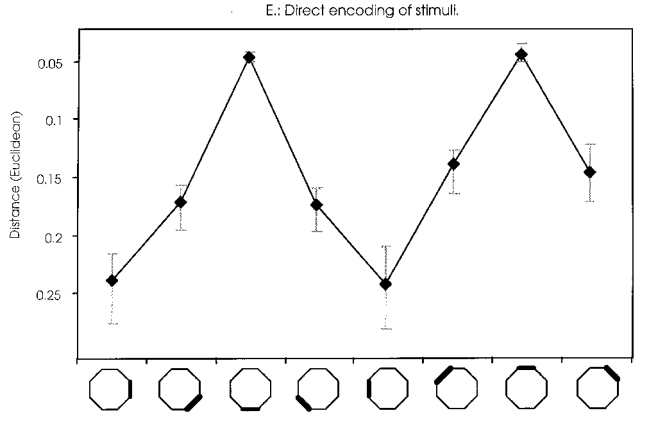


continued . . .

Figure 6. Distances between the representations for adjacent pairs of stimuli after training the model on stimuli composed of the pair of dimensions: (A) size of circle and tilt of its radial line; (B) height of a rectangle and position of embedded vertical line; (C) height and width of a rectangle; (D) length of two parallel vertical lines; and (E) arbitrarily encoded stimuli. The dimensions of (A) and (B) interact separably, those of (C) and (D) integrally. The values on the vertical axis are plotted in reverse order to ease comparison with figure 1.


Pair differ along one dimension
Pair differ along both dimensions
C.: Integral dimensions of height and width of a rectangle.
distance (Euclidean)
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
Pair differ along one dimension
Pair differ along both dimensions


D.: Integral dimensions of length of two parallel lines.
distance (Euclidean)
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
Pair differ along one dimension
Pair differ along both dimensions


E.: Direct encoding of stimuli.
Distance (Euclidean)
0.05
0.1
0.15
0.2
0.25

correlation coefficient of $r = -0.85$. It is negative since the values of the network are interstimulus distances and we postulated that distance and confusion are negatively correlated.

A two-way analysis of variance (ANOVA) was calculated on the distances between hidden unit representations over simulations and stimulus sets, which showed a significant effect of one-dimensional changes versus two-dimensional change, $F(1,150) = 6.6$, $p < 0.01$, $r = 0.21$, and of integral versus separable pairs of dimensions (*A* and *B* versus *C* and *D*) $F(1,150) = 17.4$, $p < 0.001$, $r = 0.32$. We excluded data from figure 6(E) from the analysis, since it is based on a different encoding scheme. A Pearson correlation analysis of hidden unit activation levels for each of the five different runs for each stimulus set separately showed a significant value of $r > 0.85$, meaning that over five runs the network has found a similar solution to the problem domain, i.e. relative distances between stimulus representations are highly similar.

Shepard (1964) argued that the four-cycle pattern of high one-dimensional confusion and low two-dimensional confusion, which has been observed in figure 6(A, B) as well as in the human data, cannot be explained with a Euclidean metric. Rather, a four-cornered isosimilarity contour is deemed more appropriate. In both cases —human and network—those stimuli that vary in only one dimension seem overall to be more difficult to separate than stimuli that vary in size and tilt. The dimensions of size and tilt must therefore be separable. A different picture is visible in figure 6(C, D), which contains the pairwise distances between representations of stimuli constructed from an integral pair of dimensions. In this case, there is no four-cycle pattern separating two-dimensional variation from one-dimensional variation. Instead, there is a two-cycle pattern, which matches an underlying elliptic contour. This contour is elongated with the endpoints corresponding to the two largest distances. Evidence for a difference between one-dimensional and two-dimensional variation is not available because there is no consistent alternation of distances for these variations that matches what has been observed with separable pairs of dimensions. What seems to have happened is that the integral pair of dimensions is processed holistically with the psychological dimensions out of alignment with the physical dimensions: a variation in one physical dimension will psychologically lead to a variation along both psychological dimensions.

There is, however, an artifact in Kruschke's stimuli. Scaling the two physical dimensions should not consistently place the diagonal pairs further apart in Euclidean distance in the scaled space than the aligned pairs. If we scale the rectangle height by either 2/3 or 3/2, then the distances between adjacent pairs of stimuli (starting with 2–3; diagonal pairs in cursive) change as shown in table 1.

When scaling with 2/3, the adjacent pairs 2–3 and 6–7 move closer together than the diagonal pairs 3–4 and 5–6. This is not uniform, since the diagonal pairs are actually moving closer with respect to 4–5 and 8–1. Scaling with 3/2 shows the reverse pattern. Given this artifact, it may not be reliable to use four-cycle oscillation as evidence for

Table 1. Changes in interstimulus distances after scaling.

Pairs	2–3	*3–4*	4–5	*5–6*	6–7	*7–8*	8–1	*1–2*
2/3	–0.15	–0.1	0.0	–0.1	–0.15	–0.13	0.0	–0.13
3/2	+0.22	+0.17	0.0	+0.17	+0.22	+0.22	0.0	+0.22

separable dimensional interaction, although an analysis of the topography of internal representation space of the model can reveal the separateness of dimensions.

An effect similar to integral interaction can be observed in the results from the simulations with the direct encoding (figure 6(E)). This can be explained by considering the properties of backpropagation, which are directed to extract a set of optimal representations from the stimulus space that compresses all redundant information and highlights information with important consequences. The task of the network is to produce efficient and informative stimulus representations, and as such it is biased to build the representation that most accurately reflects the correlations between stimuli and responses. In generating a set of representations no restrictions are placed on the network that preclude rotating the psychological space. This seems to have occurred with the directly encoded stimuli, suggesting that the performance of the model resembles those of human subjects confronted with stimuli constructed from integral dimensions. With the inclusion of a Gabor filter input layer, however, an additional set of restrictions has been placed on the network that prevents it from imposing a new dimension on the stimuli and instead forces it to use the extra information embedded in the encoded stimuli. Because we have direct access to the network's interstimulus distances, we can observe if the obtained data provide information about the isosimilarity contour of the standard stimulus S. The right-hand sides of figure 5 show, for each set of stimuli, the locus formed by the representations of the eight stimuli obtained from simulations with only two hidden units—although this made the task more difficult for the network, providing a success ratio of just 10%. To obtain these distances, we determined the network's representation for stimulus S (by presenting a trained network with the given stimulus) and calculated the distances between S and the other stimuli, in addition to the already available pairwise distances between the trained stimuli. We verified the distances obtained with two hidden units for similarity with the above data for three hidden units. Correlation analysis between distances for two hidden versus three hidden units proved significant ($r > 0.85$); plotting these distances for the Shepard stimuli shows a similar shape but uniformly reduced in magnitude due to the constricted activation space (figure 7). The training with one fewer hidden unit in effect implements dimensionality reduction, since the network is forced to use the largest variances to map input to output (see also Bishop 1995). Although the interstimulus distances for the two hidden unit networks are scaled with respect to three hidden unit networks, the approximation of the psychological space is accurate given the significant correlation.

It can be observed that the psychological space of the network has been warped relative to the physical stimulus space. For the separable pairs of dimensions (figure 6(A, B)), we observe that representations of stimuli which vary in only one dimension have moved closer together in psychological space and that representations of stimuli varying in both dimensions have separated. In both cases the psychological space seems to be a warped representation of physical space: although the relative positions of the stimuli are preserved—the clockwise ordering of the eight stimuli persists in psychological space—the distances between the representations have been altered as a result of learning. Critically, if we align only the primary and secondary axes of the contours with the psychological dimension axes, it directly shows that for the separable dimensions, changes involving a single physical dimension produce changes along one psychological dimension. The magnitude of such changes in psychological space remains relatively independent of other physical feature values.

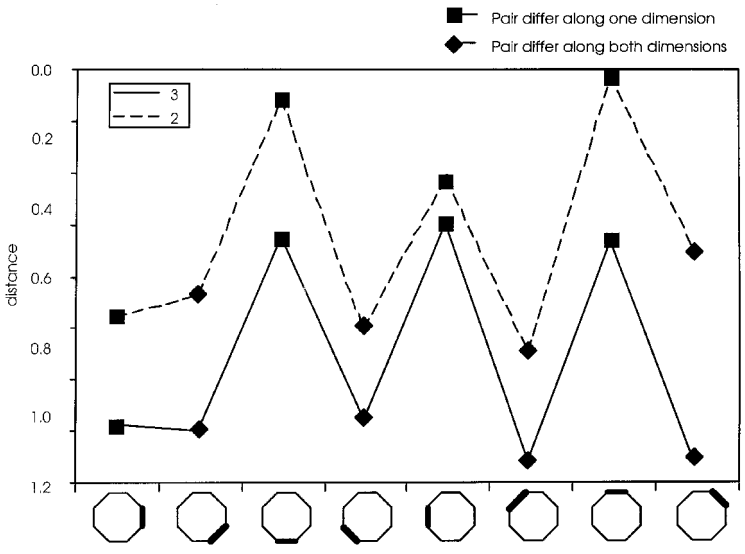


Figure 7. Interstimulus distances of hidden unit representations of adjacent pairs of Shepard stimuli for networks with two (dashed line) versus three (continuous line) hidden units. The shape of both lines is similar to the magnitude of distances for two hidden units uniformly lower than those for three hidden units (cf. Kruschke (1993)).

This warping of the physical configuration of stimuli is different for dimensions that interact in an integral manner as shown in figure 5(C, D). Alignment with psychological dimensions shows changes in one physical dimension affecting both psychological dimensions. There is also a difference in the two ellipses of both integral pairs of dimensions exhibited by the direction of the primary axis of the ellipse. The nature and direction of this axis probably reflects the salience of one psychological dimension over the other. This salience may be verbalized in the case of the integral pair of height and width of a rectangle as the salience of a dimension corresponding with the area of a rectangle.

The differences between integral and separable interaction between dimensions can be further revealed by running a hierarchical clustering algorithm (Johnson 1967, Hanson and Burr 1990), which we can use to see whether the distances between hidden unit activations contain information about the similarity structure of the represented stimulus space. Figure 8 displays the dendrograms for the hidden unit representations of the stimuli of figure 5(B, C). The similarity structure revealed in these trees indicates a remarkable difference between the two dimensional interactions. The clustering for the Kruschke stimuli show two main clusters that correspond to distances between representations of those stimuli that vary in one dimension (1 + 8 and 4 + 5) versus distances for stimuli that vary in both dimensions (2 + 3 and 6 + 7). These clusters are further subdivided into the adjacent pairwise distances. For the stimuli built from integrally interacting dimensions, no such clustering is observable. Instead, the two main clusters separate the representations of the first three stimuli from the rest, regardless of one-dimensional or two-dimensional variation. This clustering indicates that the psychological dimensions are out of alignment with the physical dimensions.

There are at least three factors that cause warping of psychological space. First, attention effects can help the subjects focus on one or more relevant dimensions

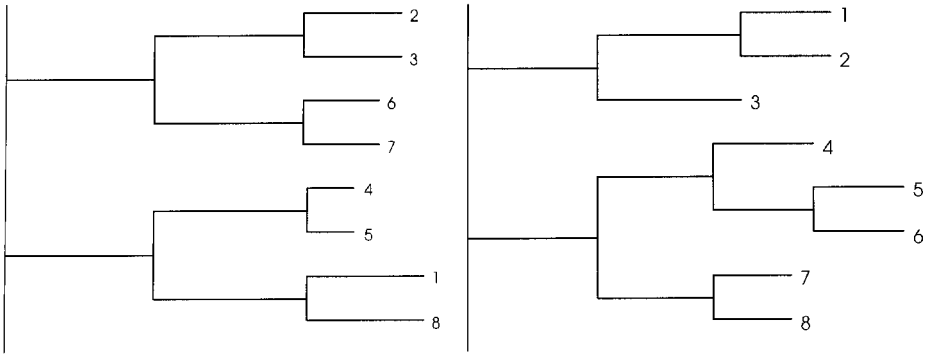


Figure 8. Hierarchical clustering tree diagrams for the hidden unit representations of the stimuli of figure 5(B) (left) and those of figure 5(C) (right). Whereas the former indicates clustering corresponding to one-dimensional versus two-dimensional variation of the physical stimuli, the latter diagram reveals lack of sensitivity to such dimensional variation.

(Goldstone 1998a). We shall visit the issue of attention in a later section. Second, there is the nature of interaction between the dimensions that compose the stimuli, as shown with the above simulation. The differences in confusion scores based on variations in either one or in both dimensions only occur with separable dimensions, since an integral pair of dimensions is perceived holistically and, as such, a variation in one physical dimension reduces to a variation in both psychological dimensions. In the case of a separable interaction, being able to attend to one dimension over the other might warp the psychological space even more, increasing distances between stimuli that vary in the dimension that receives attention (Nosofsky 1986).

Finally, one dimension can also be more salient than another one, which means that differences in this dimension are easier to discriminate than differences in the other one. In Tijsseling *et al*. (2002), both human subjects and neural nets trained to discriminate and categorize imaginary animals, which varied in the dimensions of torso radius and limb length, showed a preference for using limb differences over torso radius differences. A similar effect can be observed here. If we sum the average of opposite difference scores for all simulations of the Shepard stimuli, i.e. stimuli 5 with 8 and 1 with 4 (variations between them are variations in a single dimension of size) as well as the distances from 2 to 7 and 3 to 6 (which varied in tilt only), then the tilt dimension provides the largest interstimulus distances (see figure 9).

In Shepard (1964) a similar analysis was performed, which showed that the amount of confusion for stimuli varying in the tilt dimension is likewise lower than for stimuli varying in the diameter dimension. We cannot say for certain if the salience of the tilt dimension is an intrinsic characteristic of human visual perception or a side effect of the way the stimuli are created. Since Gabor filters are based on processing angle differences, this effect can be argued to be an artifact. However, the similarity of Gabor filters to simple cells in the primary visual cortex (Marcelja 1980, Jones and Palmer 1987) may be an indication that the same cause underlies human subject data. In the next section, we shall show that this observation matches an asymmetric interaction between the dimensions of size and tilt as reported in Potts *et al*. (1998).

We have shown that a basic connectionist model contains a potential for a simple mechanism that differentially processes integral from separable interaction between

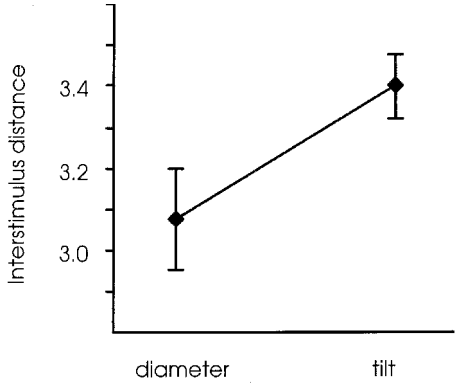


Figure 9. Sum of diametrically opposed difference scores of the network in identifying Shepard stimuli, which shows that the tilt dimension seems to be more salient. Error bars indicate minimum and maximum over five simulations (cf. Kruschke (1993)).

dimensions. But why does this work? The Gabor filter component measures certain local frequencies of the stimulus and the output of this component provides a stimulus encoding that reflects the physical properties of the stimulus. It also preserves the similarity gradient between stimuli, because the encodings are not orthogonal; rather, they reflect changes in orientation, location and angle between the various stimuli (Goldstone *et al*. 1996). Figure 10 captures this similarity gradient. It is an illustration of how, for the stimuli that vary in separable dimensions of size and tilt and for the stimuli that vary in integral dimensions of height and width of a rectangle, the Gabor filter model processes the physical information and produces an output that shows the relational changes between the stimuli. For example, figure 10(A) shows how the response for an angle orientation of 90° gradually increases from stimulus 1 to 4 and then gradually decreases again. Based on this orientation information, stimulus 1 would be more similar to stimulus 2 than to stimulus 3.

The relevance of preserving the similarity gradient has been shown by Harnad *et al*. (1995) and Tijsseling and Harnad (1997). Using a backpropagation network, they showed that stimulus encodings that preserve the similarity gradient (also called iconicity) are a necessary property of categorical perception models (Harnad 1987a). It is, however, not enough just to encode the similarity gradient of any set of stimuli, since this fails to capture the differences between sets of stimuli that have a comparable similarity gradient, but a different physical structure. Three line segments of 3, 4 and 5 cm can be encoded in a way that captures their iconicity by using a thermometer coding converting the stimuli to three eight-bit vectors: 11100000, 11110000, 11111000. This encoding captures the fact that a line segment 3 cm long is more similar to that which is 4 cm long than to a line 5 cm long, but it does not reflect the physical structure of the set, characterized by the relative position and angle of the lines.

To explore further the relative contribution of the Gabor filter, we performed multidimensional scaling of the Gabor filter responses. The results are displayed in figure 11 and show that the Gabor space topography has captured the relative similarities between the stimuli. Stimuli with a separable pair of dimensions are observed to be closer in Gabor space if they differ in only one dimension (figure 11(A, B)). This is not as strong with integral pairs of dimensions (figure 11(D)), or even absent (figure 11(C)). This suggests that the perceived integration or separation

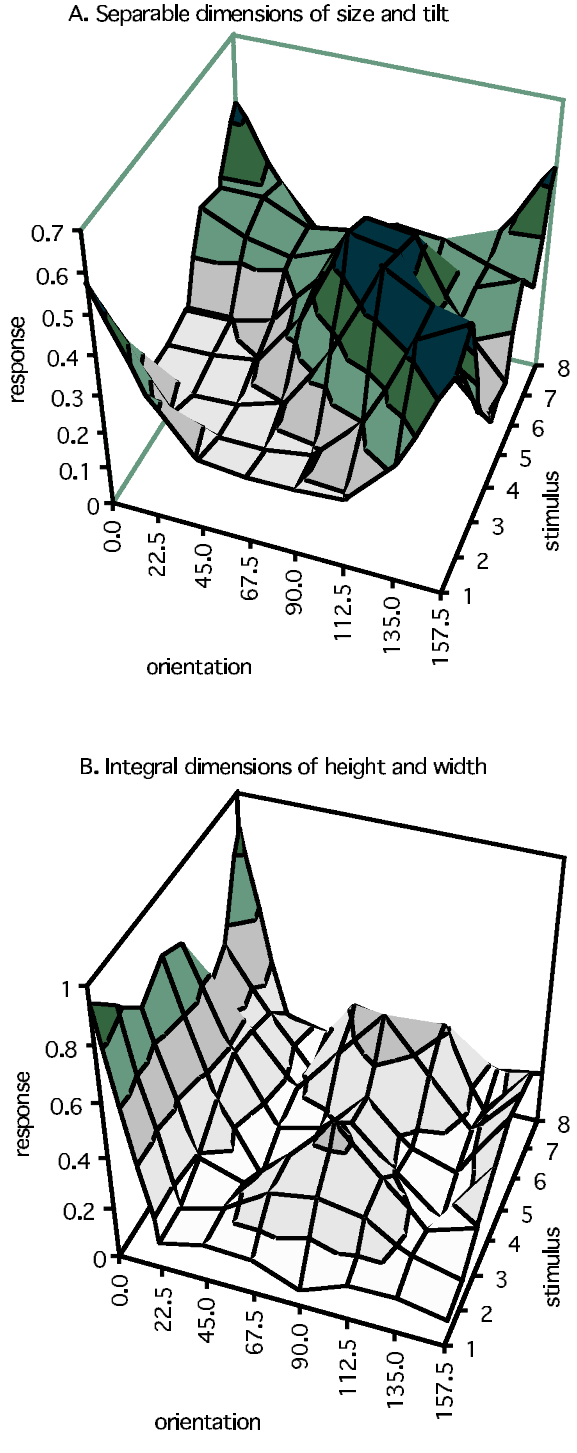


Figure 10. Three-dimensional display of the Gabor responses at eight orientations for a spatial frequency of 3.5 and at the centre of the image for the stimuli with (A) separable dimensions of size and tilt and (B) integral dimensions of height and width of a rectangle.

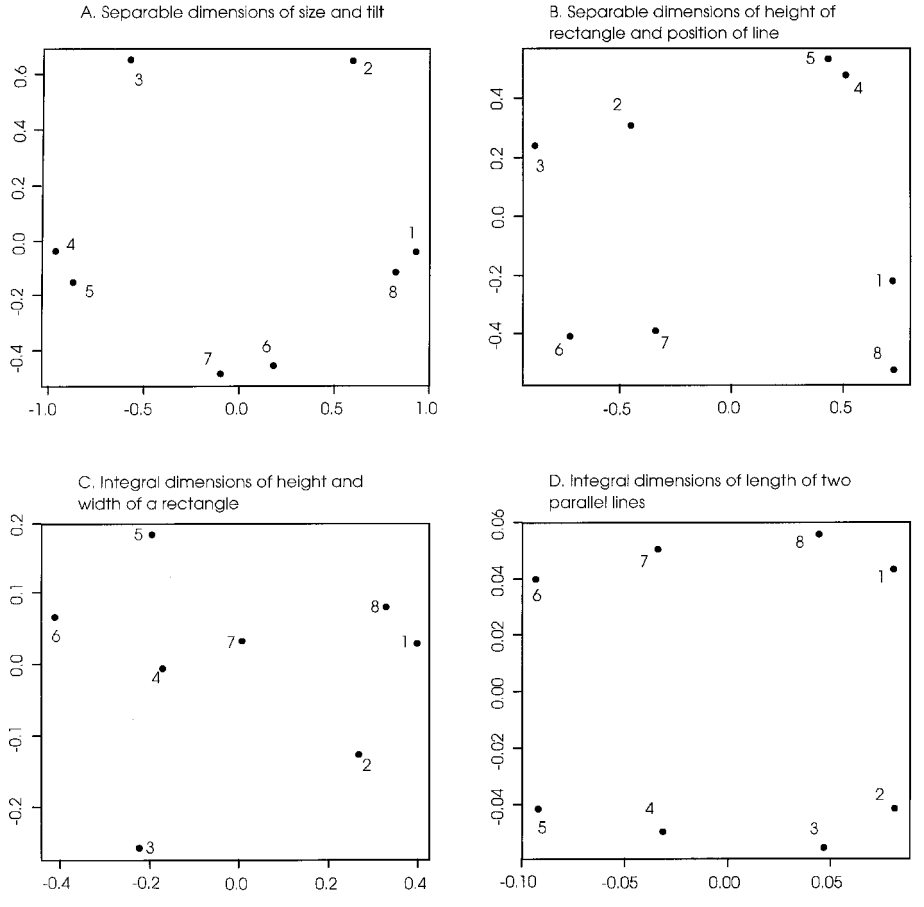


Figure 11. Multidimensional scaling of the Gabor space topography for the four stimulus sets.

of stimulus dimensions may be, primarily, a natural result of Gabor-like sensory processing in the primary visual cortex (Marcelja 1980).

The richness of information that Gabor filter processing of the stimuli produces indicates that basic properties of the visual system (in this case modelled with crude Gabor filters) may already capture most of the relevant differences between integral and separable interaction between dimensions. It is a widely held view (Gibson 1991, Goldstone 1998a) that dimensional integration in perception naturally precedes the learning of labels for stimuli. Indeed, dimensional integration effects arise early in perceptual processing before substantial cognitive experience with stimuli is built up, a view consistent with development of perceptual learning in children (Kemler and Smith 1978, Smith and Kemler 1978) and effects of expertise with stimuli (Burns and Shepp 1988).

Given the topography of the sensory information provided by Gabor filters, the role of the identification network is to reduce identification, discrimination and categorization error by applying redundancy compression and predictive differentiation on the responses of the Gabor filter component (Gluck and Myers 1993). The preserved similarity gradient provides the model with sufficient information to enhance relevant

features and suppress irrelevant information. This process of compression and differentiation causes the observed warping of internal representations. Although the Gabor filter component captures dimensional interaction, it may not be sufficient to produce correct identification of stimuli. The effect of warping may be an explanation for the linear change in confusion scores with every other adjacent pair of stimuli shown in figure 3. The identification learning process by the human subjects may have separated representations further in order to reduce potential identification mistakes.

A Gabor filter as a crude sensory processing component in combination with a backpropagation network for redundancy compression and predictive differentiation provides a mechanism that is basic and simplified, but powerful enough not only differentially to process the interaction between dimensions but also to use this perceptual information for cognitive processes of identification, discrimination and categorization. In the next section the model is applied to Garner classification tasks to explore further the capacities of the model in explaining dimensional interaction and its possible implications.

6. Applying the model to Garner classification tasks

6.1. *Introduction to Garner classification tasks*

We have shown that the Shepard paradigm, which consists of identifying stimuli that are organized in an octagonal configuration, provides a method for determining the interaction between the two dimensions that make up these stimuli. The distribution in psychological space of representations of stimuli composed of a separable pair of dimensions retained the octagonal configuration of its physical counterpart, although it may be warped differently for stimuli composed of different separable pairs of dimensions. With integral pairs of dimensions, this distribution of representations corresponded to an elliptic locus, a transformation from physical space that indicated that the two dimensions were perceived holistically.

To show that these results are generalizable and consistent, we employed the same stimuli but using a different classification task, which is described in Potts *et al.* (1998). In this task, subjects had to classify stimuli into two categories. For each condition, only one dimension was relevant to the sorting task: subjects were asked to pay attention to the relevant dimension and sort the stimuli according to the value on that dimension. Garner and Felfoldy (1970) found that for integral pairs of dimensions, the ease with which subjects could complete this task, as measured by the time it took them to complete it, was critically dependent on the values taken by the irrelevant dimension. For separable pairs of dimensions, reaction times were unaffected by manipulations to the irrelevant dimension.

Although Garner and Felfoldy (1970) gave a good indication of how subjects process separable versus integral pairs of dimensions, several researchers (e.g. Pomerantz and Garner 1973, Ward 1982, Potts *et al.* 1998) conducted experiments where their own data were less conclusive about a strict dichotomy between integral and separable. For example, Potts *et al.* (1998) showed how for dimensions of circle-diameter and radial line-angle, variations in the location of the stimuli in the two-dimensional physical space can produce different forms of interaction between the dimensions. They argue that since the radial line of a circle provides information about the size of its circle, but size itself does not provide information about the tilt, there is an asymmetric integrality between these two dimensions. In other words, since a

radial line indicates the size of a circle (a longer radius means a bigger circle), there is a weak form of integral interaction between size and tilt, which can become evident under certain conditions (Potts *et al*. 1998). These conditions are illustrated below.

Potts *et al*. (1998) had subjects classify stimuli into two categories. The stimulus sets varied in the difficulty of discrimination between the two sizes and angles, and whether the radial line touched the perimeter of the circle. Three examples are illustrated in figure 12. In figure 12(A), stimuli were circles with a diameter of 41 mm and 63 mm with a tilt of 12° clockwise or counterclockwise. In figure 12(B), discrimination was decreased by reducing the tilt to 11° and the diameter to either 45 or 55 mm. Finally, in figure 12(C), the weak interaction between tilt and size was removed by fixing the length of the radial line at 32 mm and, therefore, preventing it from touching the perimeter of the circle.

There were five different classification tasks, two single dimension tasks, two correlated dimension tasks and one orthogonal dimension task. In Single Dimension 1, either the size varied and tilt was kept at a clockwise position, or tilt varied and size was large. Single Dimension 2 had either size varying and tilt counterclockwise or tilt varying with small circles, providing two conditions. Correlated dimensions were either positive or negative: large and clockwise versus small and counterclockwise or large and counterclockwise versus small and clockwise, respectively. Finally, in the orthogonal task both dimensions varied orthogonally, with either size or tilt designated as the relevant dimension.

The reaction times, illustrated in figure 13, reveal a striking variation across stimulus sets and across dimensions. For example, in figure 13(A) an asymmetric interaction can be observed: correlating both dimensions did not produce any facilitation compared to a single dimension variation. However, varying size did interfere with tilt when the latter was the relevant dimension. A different pattern of results, shown in figure 13(B), was obtained when size was made less discriminable (see figure 12(B)): variation in size produced slower reaction times. Decreasing discriminability of the size dimension, therefore, weakened the asymmetric interaction, although this asymmetry was not reversed. Potts *et al*. (1998) observed that subjects still found it more difficult to process tilt when size was small than when size was large. They suggest that the system's preference for one dimension over the other may reflect more than a preference for superior discriminability (Potts *et al*. 1998: 108), i.e. it could be

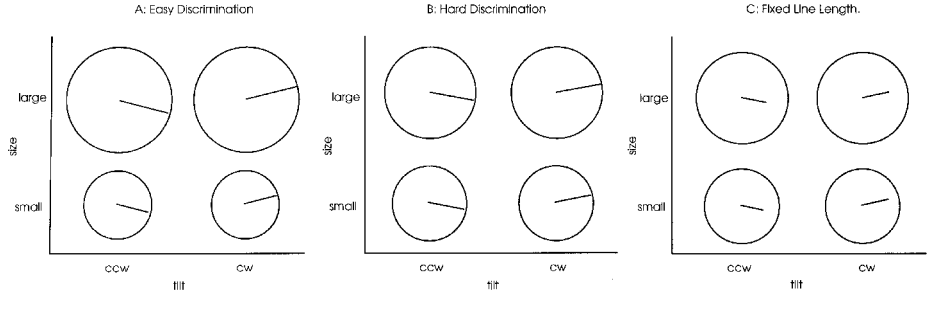


Figure 12. Stimulus sets for experiments by Potts *et al*. (1998). (A) Circle sizes were 41 or 63 mm, with tilt being 15° clockwise or counterclockwise (easy discrimination). (B) Circle sizes were made less discriminating by reducing them to 45 and 55 mm. Tilt was decreased to 11° (hard discrimination). (C) Stimuli were as in (B), but the length of the radial line was fixed at 32 mm (increased separability).

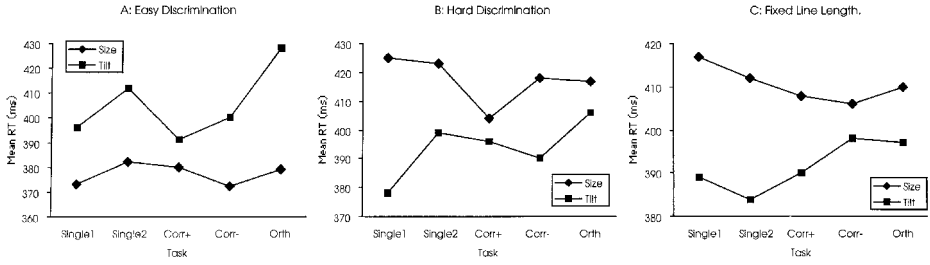


Figure 13. Mean reaction times for tilt and size dimensions as a function of task. The classification tasks are explained in the text. (A)–(C) show the reaction times for the corresponding figures of figure 13. Figure reproduced from Potts *et al*. (1998). Error bars were not provided in the original figure.

factors determined by the nature of the sensory system. In the previous simulations, we showed that there was a preference for the tilt dimension, an observation also indicated by Shepard (1964). It was suggested that this preference might be a natural artifact of the Gabor filters.

If this asymmetric interaction is removed by keeping the length of the radial line constant at a certain value (see figure 13(C)), then the reaction times indicate complete separability of dimensions because the radial line does not supply information about the size of the circle anymore. Although the dimension of tilt produced faster reaction times in figure 13, there was a lack of reaction time differences as a function of task within each dimension (Potts *et al*. 1998).

6.2. *Simulations with the demonstrative model*

Since our focus is not on replicating the results from Potts *et al*. (1998), but on exploring the correlation of data from this Garner paradigm with data from the Shepard paradigm, we did a simulation of the above procedures using four stimuli in the range of dimensional values used in our Shepard simulation (see figure 14(A)). We measured the number of epochs it took the model to reach a pre-specified criterion (*MSE* of 0.1) as a substitute for reaction time.[3] We assume that the relative ease with which the network separates the representations of stimuli correlates with the speed with which a network identifies a stimulus. Since the model does not have the ability to attend specifically to one dimension, we were not able to implement both versions of the correlated task. In other words, when sorting a large circle with a clockwise tilt from a small circle with a counterclockwise tilt, it is not possible to instruct the model to sort these two stimuli according to either size or tilt. The model was instead taught just to sort the stimuli.

The architecture of the model differed from the previous simulation in that the output layer had one extra unit to encode category membership (set as zero for the first category and one for the second category). The network's task, therefore, is to learn both to identify the stimulus item with a unique label as well as to categorize it. We ran 10 simulations for each task, using different initial random weights from the range [–1.0, 1.0], using the same weight set for all tasks. The reason for this is to ensure that differences in the model's performance for each task cannot be attributed to a difference in initial weights. As such, differences in performance reflect the difficulty of the task, not the configuration of weights.

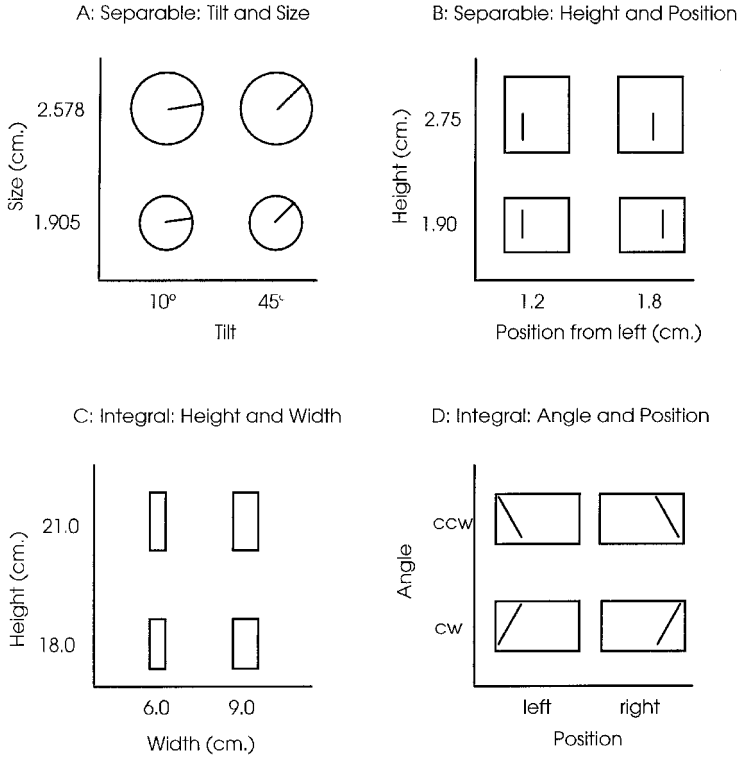
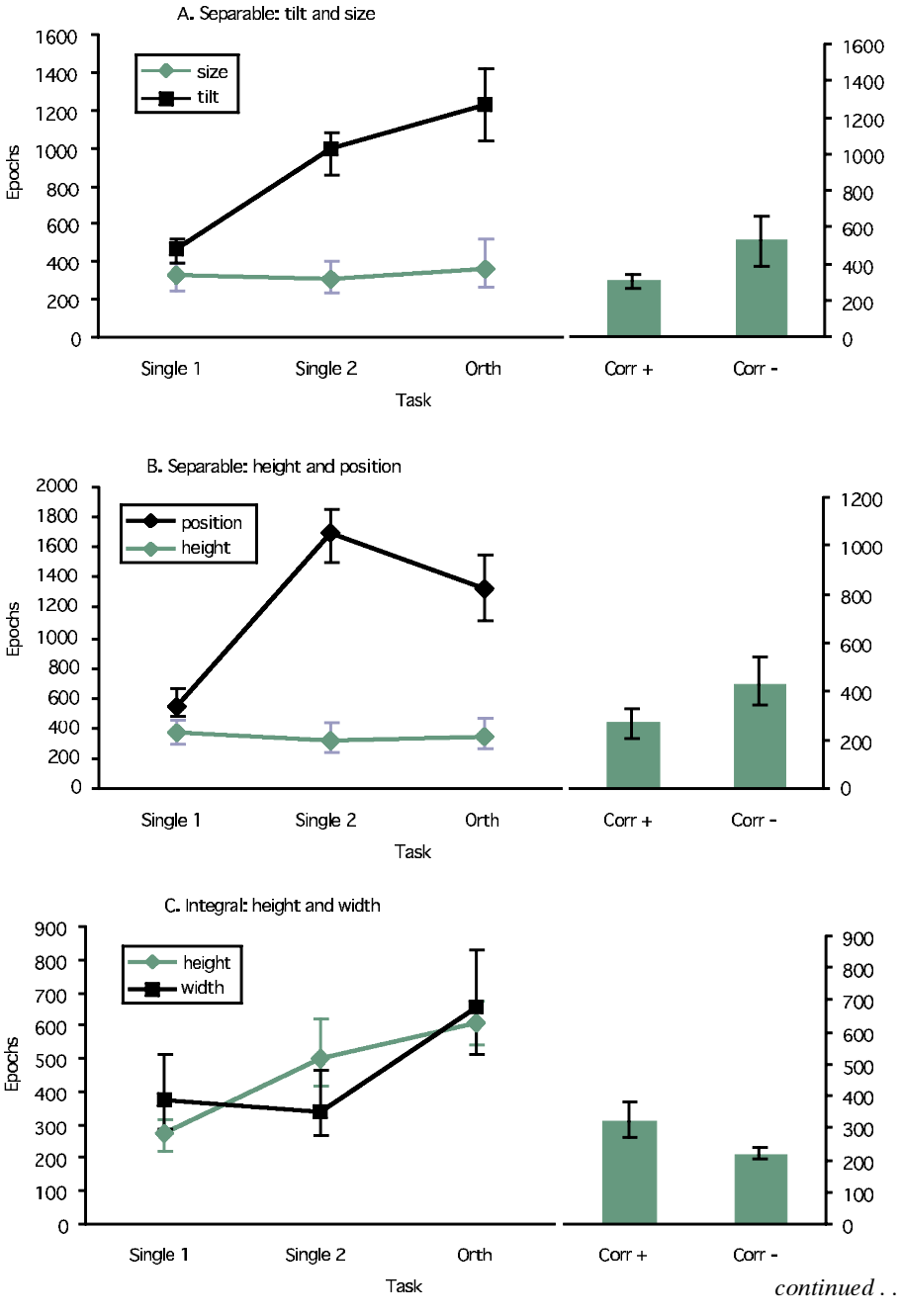


Figure 14. Stimuli used in the simulations of the Garner speed sorting task. The values on each dimension are in the range of the values used for the Shepard task in order to be able to relate the results obtained here with those obtained with the Shepard task. In each figure, stimuli are composed of: (A) separable pair of dimensions of diameter of circle and angle of radial line (Shepard 1964); (B) separable pair of dimensions of height of a rectangle and position of an embedded vertical line (Kruschke 1993); (C) integral pair of dimensions of height and width of a rectangle (Monahan and Lockhead 1977); and (D) integral pair of dimensions of location and angle of a straight line (Redding and Tharp 1981).

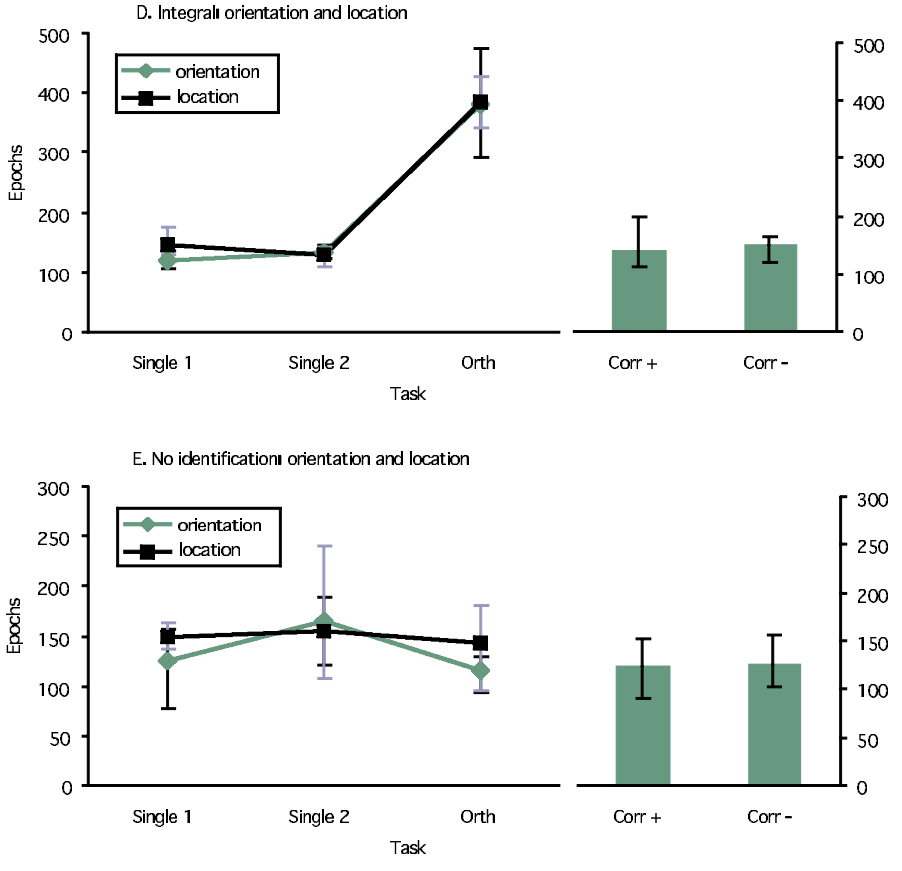
The results, illustrated in figure 15(A), show the same kind of asymmetric interaction as in figure 13(A). Variations in the tilt dimension required more epochs to sort, and there was interference from the size dimension when this was varied orthogonally. When the dimension of size was relevant, there was no interference from the tilt dimension in all tasks. When both dimensions varied, performance on the correlation task was dependent on the respective values stimuli took along their dimensions, which is also evident in data from Potts *et al.* (1998). This asymmetry of the interaction between the dimensions of size and tilt confirms what we observed in the results of the Shepard task. In this task, both the confusion scores and the interstimulus distances varied with each adjacent pair of stimuli. In particular, the confusion scores were higher (and the interstimulus distances lower) for adjacent stimuli that had small circle sizes than adjacent stimuli that had large circle sizes.

Figures 15(B–D) display the results of simulating the above task with the other kinds of stimuli used with the Shepard task. These stimuli are illustrated in figure 14(B, C). Figure 14(D) shows an additional set of stimuli that were employed in experiments by Redding and Tharp (1981). The dimensions that compose these



continued . . .

Figure 15. Results of the speeded sorting tasks with the stimuli of figure 13. (A)–(D) correspond with figure 14(A)–(D), and so forth. (E) The number of epochs it took a plain feedforward network to sort straight lines that vary in orientation and location (Redding and Tharp 1981). The results for the correlated task are displayed in a separate graph, because although the model is capable of perceiving the dimensions that compose a stimulus, it cannot be instructed to attend to a specific dimension. Since both dimensions varied for the two stimuli of the correlated task, we cannot infer which dimension was used by the network to learn the task or whether both were relevant.



stimuli are location and orientation of a single straight line and they have proven to be integrally interacting (Redding and Tharp 1981). In figure 15(B), it can be observed that the dimensions of height of a rectangle and position of an embedded vertical line, which were used as interacting dimensions by Kruschke (1993), show asymmetric interaction: irrelevant variation of position had no effect on sorting time of height differences, but variation in height affected sorting stimuli on position. As for the integral interacting dimensions shown in figure 15(C, D), in both cases there is a clear effect of interference from the irrelevant dimension in the orthogonal sorting tasks, with the dimensions of Redding and Tharp (1981) showing a near perfect symmetry of interaction.

We argued in the previous section that the Gabor filter component captures most of the relevant difference between integrated and separated dimensions, the perceptual representations of which are further separated by the backpropagation component under task constraints such as identification learning. In this second set of simulations, categorization task performance is shown to be dependent on the structure of the perceptual representations from the Gabor filter component. The Gabor responses for separable pairs of dimensions allow categorization based on a single dimension to be learned with relatively little interference from the other varying dimension, but with integral pairs interference from irrelevant dimensions is relatively stronger.

The magnitude of this cross-dimensional interference, however, arises only when the networks are required to identify the individual stimulus items at the output layer. When the task of the network is only to categorize (i.e. there is no identification process), there is no evidence of dimensional interference. Figure 15(E) shows the results from simulations with a 'categorization only' backpropagation network (e.g. with only one output node). When this network was presented with Gabor filtered responses from the Redding and Tharp stimuli, the number of epochs it took to sort the stimuli for each task did not correspond with an integral interaction between dimensions. Rather, the 'categorization only' network quickly attended to only the relevant dimension. The reason for this pattern of results is that in the absence of identification training, the network does not have to separate individual stimuli to lower an identification error, since there is no such error signal. It is only given a categorization task, and the stimulus representations are consequently clustered according to the categories they belong to, ignoring any interstimulus variations except those that are relevant for placing a reliable category boundary (see figure 16).

This pattern of results implies that human participants covertly identify presented stimulus items, a process that forces attention to be directed to both dimensions when both are needed to discriminate the stimuli. This in turn has a profound effect on the categorization performance. The tendency to identify stimuli may have been a result of verbally provided information on the experiment by the controller, asking the subject to focus on a dimension. It is also possible that sorting a set of stimuli implies the requirement to discriminate. More interestingly, it suggests that categorization and identification are intrinsically linked in the cognitive system. Such notions are basic to categorization theories such as exemplar models, which assume the automatic retrieval of exemplars in response to stimulus presentation (Medin and Smith 1984, Nosofsky 1992).

In one important aspect the described model failed to fit the behavioural data on category learning. Human subjects can be instructed to attend to one single dimension in a category-learning task. For example, when sorting stimuli in which variation in size correlated with variation in tilt, subjects can be asked to sort the stimuli according to size and ignoring variance on the other dimension. The model does not have such an attention mechanism and will therefore use variance on both dimensions in successfully sorting the stimuli. The next section describes a possible mechanism to introduce attention in a backpropagation model.

7. Providing a mechanism for selective attention

7.1. *Introduction*

The model's lack of employment of any kind of attention in processing multi-dimensional stimuli might make it appear inferior to models that capture dimensional interactions only through the fitting of experimental data. For example, in ALCOVE, each input node encodes a dimension or feature and the activation of all these nodes is gated by multiplicative attention strengths (Kruschke 1992). The advantage of ALCOVE is that one dimension can be ignored by reducing the attentional strength of the corresponding input node. Successful theories of categorization need a notion of selective attention because adaptation often occurs through the increase of attention to dimensions and features that are important for the categorization task and/or the decrease of attention to irrelevant dimensions and features (Nosofsky 1986, Goldstone 1998b). A well-known paradigm used to show how human subjects learn

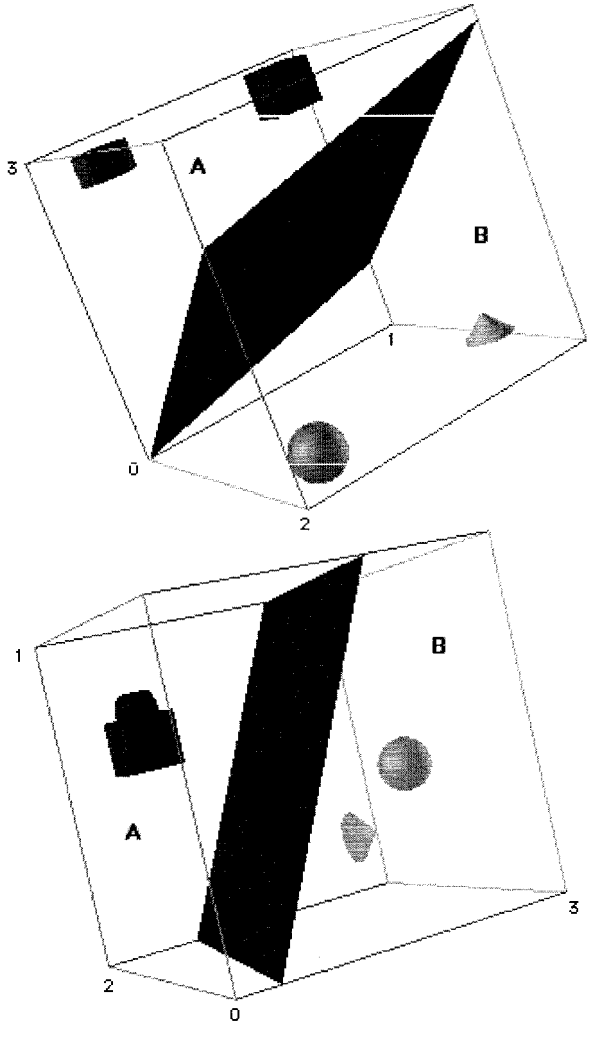


Figure 16. Location of representations of the Redding and Tharpe (1981) stimuli in hidden unit activation space of the identification + categorization network (left) and the categorization-only network (right). The three axes measure the corresponding activation level of each of the three hidden units 1 to 3, and each category space is indicated by the letters A and B. Training the network on identification has increased the within-category distances as well, compared with the categorization-only network. Note the closeness of the category A instances in the categorization-only hidden unit activation space; identifying and discriminating each of these two instances would consequently be nearly impossible.

to pay attention to a particular dimension of a presented stimulus is a filtration versus condensation task (Posner 1964, Garner 1974, Gottwald and Garner 1975, Potts *et al.* 1998).

In a variation by Kruschke (1991, 1992), subjects had to learn to categorize eight stimuli that varied along two dimensions. The stimuli were rectangles that had a vertical line embedded in them. The height of the rectangles and the position of the vertical line varied (figure 17(A); see also figure 5(B)). There were four groups of subjects and each of these groups had to learn a different task (figure 17(B)): two

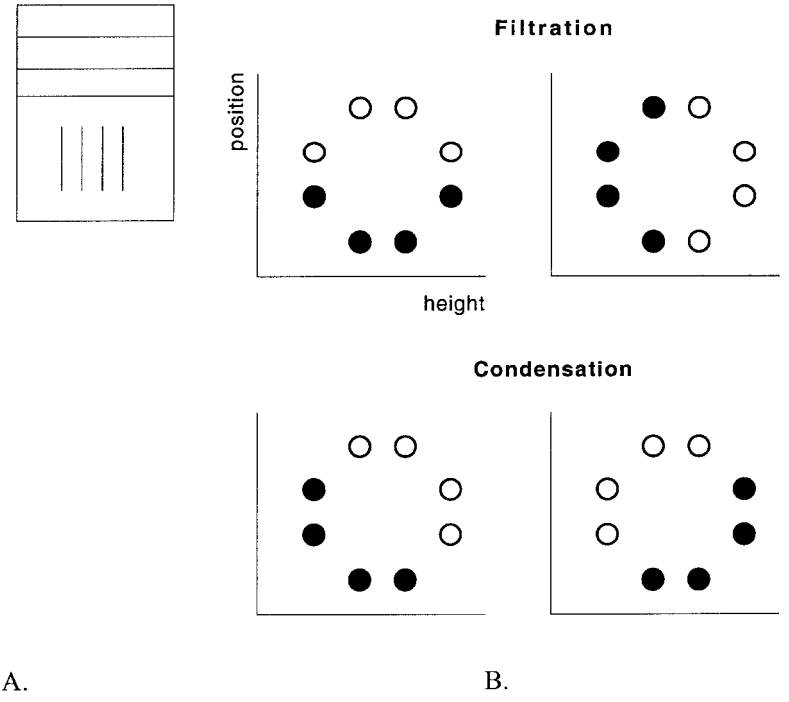


Figure 17. (A) Illustration of the stimuli used by Kruschke (1991). This figure shows the superimposition of all four possible heights of the rectangle and all four possible positions of the embedded vertical line. (After Kruschke 1991.) (B) The four category structures used in Kruschke's (1991) experiment. The horizontal axis corresponds to the dimension of height of the rectangle in figure 13(B), and the vertical axis to the dimension of position of embedded vertical line. The top panel displays two filtration tasks and the bottom one condensation tasks.

tasks involved sorting the stimuli into two equal-sized categories along one dimension only, and two tasks involved sorting along both dimensions. Human subjects consistently learned tasks that involved one relevant dimension only (filtration) statistically better than tasks that involved both dimensions (condensation). See figure 18(A) for an illustration of this difference in learning performance.

This aspect of human categorization is difficult to explain using the standard backpropagation neural network (Rumelhart *et al*. 1986). In backpropagation, the weights being adapted create hyperplanes that carve up the activation space of the network. The orientation of an arbitrary hyperplane can be in any direction in this space. In this respect, category boundaries can be placed between any two groups of inputs, which means that there would be no performance benefit for a filtration task over a condensation task. Learning condensation tasks would actually be faster for a standard backpropagation network because the extra information available from variance along the other dimension would make it easier to separate hidden unit representations (Kruschke 1993).

Kruschke (1991, 1992, 1993) provided a neural network model that can simulate human behaviour in a filtration versus condensation task. His model, ALCOVE, is a feedforward network with three layers of nodes. The input nodes encode the stimulus, with one node per psychological dimension, such as colour or shape. Every single input node has an attention strength associated with it, which reflects the relevance

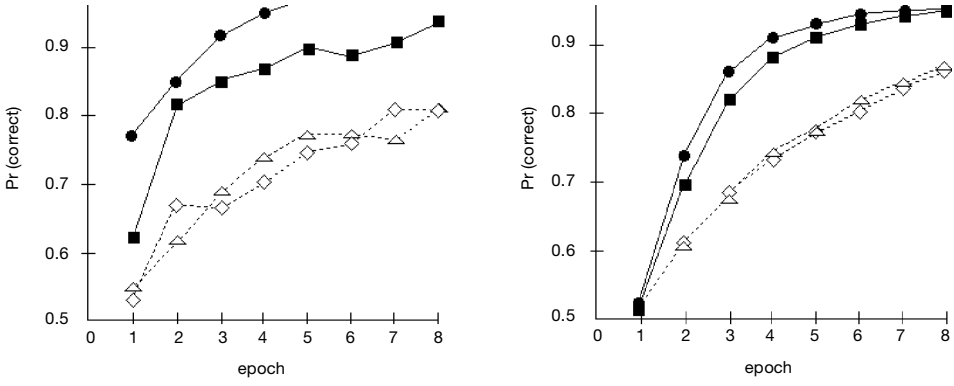


Figure 18. Panel A shows human learning data and panel B shows the simulation results with ALCOVE (Kruschke 1991). Each datum shows the percentage of correct responses for one learning block. (•) Filtration, position relevant. (■) Filtration, height relevant. (△) Condensation, category boundary along the left diagonal. (◇) Condensation, right diagonal boundary. (Reprinted from Kruschke 1991.)

of the corresponding dimension for the categorization task. Each node in the hidden layer represents a training exemplar. The activation of a hidden node reflects the similarity of the presented input to the exemplar that is represented by it. The output layer itself consists of nodes that correspond to categories to be learned. Figure 18(B) shows that this model shows filtration advantage over condensation.

Although ALCOVE models human learning data well it has its limitations. One needs to specify the psychological dimensions of inputs, which makes the model difficult to apply to real inputs such as images. Tijsseling and Harnad (1997) and Tijsseling (1998) argue that categorization models should develop their own psychological representation of the input space in order to explain categorical perception (Harnad 1987b) and as a solution for symbol grounding (Harnad *et al.* 1995). ALCOVE, however, is a psychological learning model devised to fit and explain specific human experimental data. In this paper, on the other hand, we want to do away with the *post hoc* determination of relevant dimensions and the need to incorporate data from subjects' similarity ratings in order to determine if dimensions interact separably or integrally. Instead, our approach argues for models that aim to provide a mechanism for both the perceptual and associative processes in category learning. A successful category model should try to solve the question of how the psychological space is related to physical space and, subsequently, how this psychological space is carved up into the necessary categories based on environmental feedback.

Therefore, what is needed is to incorporate selective attention in the model. We shall derive this mechanism from Kruschke's single-layer learning model, called ADIT (Kruschke 1996). This model is similar to ALCOVE as it also uses dimension values as inputs and attaches an attention strength to each of these dimension values. The difference is that ADIT is based mainly on Mackintosh's well-founded theory of selective attention (Mackintosh 1975). We applied the principles underlying the attention mechanism of ADIT to backpropagation. In the next section, we shall describe this attention mechanism and its performance in simulating human behaviour in a filtration versus condensation task.

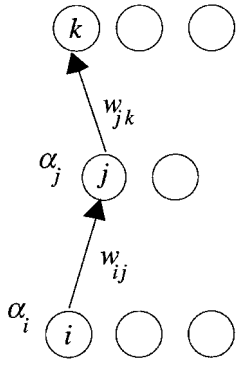


Figure 19. Illustration of an attention-enhanced backpropagation network. The bottom layer of nodes is the input layer; the middle is the hidden layer and the top layer contains the output nodes. Each input node and hidden layer has an attention strength of α_i. Connections between nodes i and j have a weight w_{ij}.

7.2. *Implementing an attention mechanism in backpropagation*

The backpropagation network is modified such that each input node and hidden unit is supplied with an attention strength α, which determines the importance of a particular stimulus feature for the corresponding response. Clearly, output nodes do not have attention strengths associated with them. Figure 19 displays a graphical representation of an attention-enhanced backpropagation network. In addition to associating attention strengths with units, the learning procedure of backpropagation is modified as well: before the weights are updated, attention is first shifted to relevant information in the current input based on the errors calculated at the output layer.

Derivation of the attention shifting algorithm is described in Appendix B. Similar to the weight adaptation in backpropagation, attention shifting is also a gradient descent method: attention should be shifted to those input values that reduce the error at the output layer the most. On presentation of an input, all input nodes and hidden units are assigned similar normalized attention strengths, according to:

$$\alpha_i = N^{-1/\sigma}$$

in which σ is a freely estimated parameter ($\sigma > 0$) and N is the number of nodes. This normalization causes the length of the attention vector to equal 1.0 when measured using a Minkowski r metric where $r = \sigma$:

$$\left(\sum_i \alpha_i^{\sigma} \right)^{1/\sigma} = 1.0.$$

After activation is propagated from the input to output layer, the errors are calculated and used to adapt the attention strengths of hidden units and input nodes, respectively. If a new attention value is negative, it is set to 0.0. The new attention strengths are normalized again, according to:

$$\alpha_i = \frac{\alpha_i}{\sigma\sqrt{\sum_j \alpha_j^{\sigma}}}.$$

After adaptation, activation is repropagated from the input layer to the output layer using the new attention strengths. The errors are then recalculated and weights are changed accordingly. This procedure repeats for every input pattern.

7.3. *Simulating Kruschke's filtration versus condensation task*

The attention-enhanced network was tested with Kruschke's filtration versus condensation task (Kruschke 1991). In a filtration task only one dimension is relevant, whereas in a condensation task both dimensions are relevant. It has been found that subjects find filtration tasks easier to learn because they can attend to one dimension only (Kruschke 1991). As argued earlier, standard backpropagation does not show this performance discrepancy because the nature of its weight adaptation algorithm allows category boundaries to be placed along arbitrary directions.

The experimental set-up of the model was identical to the previous simulations. Images were preprocessed with Gabor filters tuned to several different angles. The output layer carried one extra unit to encode category membership (the target was zero for one category and one for the other) and the hidden layer contained five instead of three hidden units because addition of attention weights made learning relatively more complicated. Simulations with both the previous and attention-enhanced model were run until the mean-squared error dropped below a criterion of 0.1. The learning rate η was set at 0.5, and the attention parameters λ_α and σ were set to 15.0 and 0.05, respectively. The range for the random initial learning weights was between –1.0 and 1.0.

The learning curves for standard and attention-enhanced backpropagation are shown in figure 20. Attention-enhanced backpropagation (panel B) not only processes filtration tasks more accurately than standard backpropagation (panel A), but also learns faster overall. A one-way ANOVA to compare condensation with filtration performance revealed a significant interaction ($F(1,14) = 30.1$, $p < 0.0001$, $r = 0.83$). The benefit of selective attention is that irrelevant information in the input stream is ignored during weight adaptation. In this sense, learning a filtration task becomes easier for the network, since the lack of irrelevant information does not hinder category boundary placement anymore.

Figure 20(B) also shows an improvement on Kruschke's own simulation results. The human data of figure 18(A) show that during the first training block there is already improved filtration advantage over condensation. This initial difference is not captured in Kruschke's results (figure 18(B)), but it is exhibited by the attention-enhanced backpropagation model. The latter displays a quick attention shifting potential, which benefits from adaptability to changing information in the input streams. The cause of this advantage is the implementation of an attention-change mechanism which provides novelty-based learning: in the first block of training everything is novel, which causes relatively faster learning compared to learning in later blocks, when information has already been presented. This mechanism might reflect what is happening psychologically, but it needs more evidence to suggest a critical factor in early-block learning.

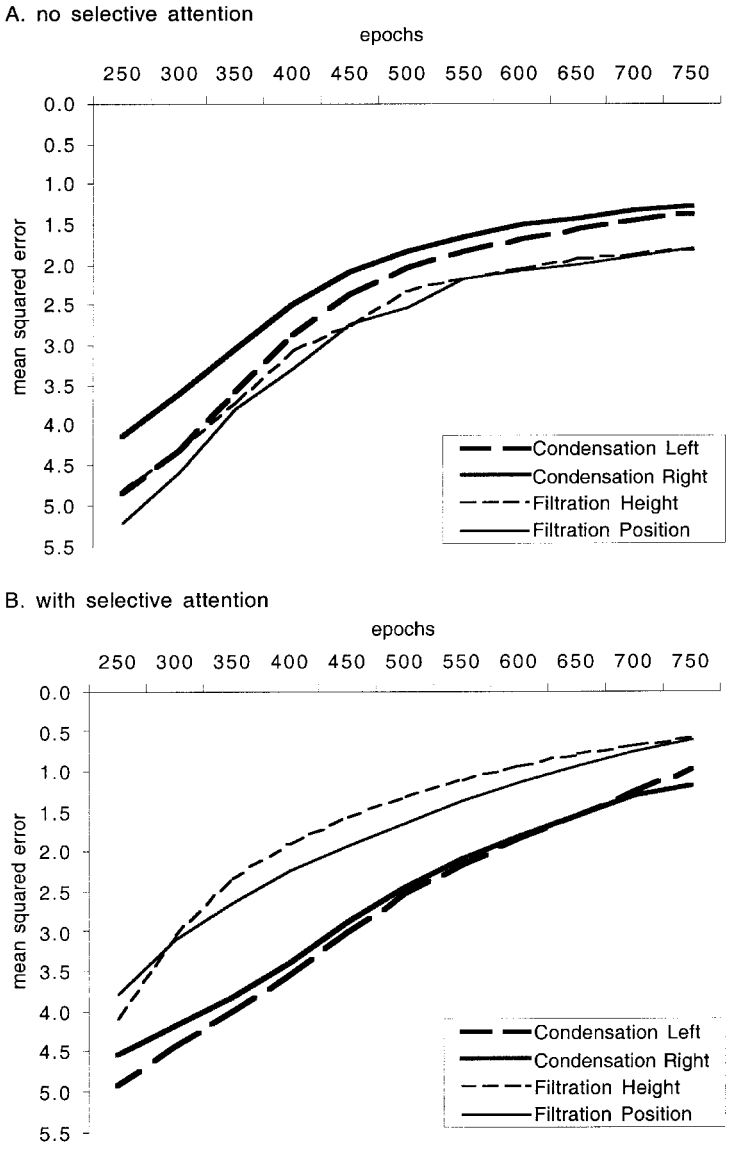


Figure 20. Category learning data from the filtration versus condensation task for standard backpropagation (panel A) and attention-enhanced backpropagation (panel B). The vertical axis corresponds to the mean-squared error and the horizontal axis correspond to the number of epochs. Not only does attention-enhanced backpropagation find filtration tasks easier to learn, it learns all tasks roughly twice as fast as standard backpropagation.

8. Discussion and conclusion

We have presented a mechanistic cognitive explanation of the problem of dimensional interaction. Prior approaches to the issue of separable versus integral pairs of dimensions have been concerned with conducting experiments with human subjects and devising statistical models to fit experimental data. Although these approaches yield useful data, they do not clarify what mechanism underlies the transformation from

a raw stimulus to an internal representation and how this transformation affects the interaction between a pair of dimensions. We argue that connectionist models supplemented with a sensory-preprocessing component naturally possess the capacity to differentiate between the various interactions between dimensions. As such, the analysis of the way these models process sensory stimuli can aid researchers not only with determining how it might be done in the human cognitive system, but also in interpreting human subject data.

We have supported our hypothesis using demonstrative simulations with a simple backpropagation 'identification network' combined with a Gabor filter input layer to filter the sensory input. The use of Gabor filters (Gabor 1946, Daugman 1988) to encode monochromatic stimulus images provided a way to preserve not only the similarity gradient of the stimuli, but also their physical properties. Indeed, by analysing the Gabor filter responses, we observed that this component captured most of the relevant difference between integrated and separable pairs of dimension, from which we inferred that the perceived dimensional interaction could already be embedded in the Gabor-like sensory processing of the primary visual cortex. Consequently, the Gabor filter component can predict the dimensional interaction of a given set of stimuli by examining the results of multidimensional scaling. Similarly, in Tijsseling *et al*. (2002), both human subjects and self-organizing competitive neural nets were trained to discriminate and categorize two-dimensional images of three-dimensional shaped imaginary animals, which varied in the dimensions of torso radius and limb length. For the networks these images were preprocessed with 100 overlapping Gabor filters, spread over the entire image. A multivariate analysis of the responses from these filters revealed dimensional differences and a salience of the limb dimension. This preference for limb differences was proven to be persistent for both networks and human subjects and over several identification and categorization tasks.

The initial perceptual representation or the structure of the Gabor space can be further 'warped' by cognitive processing, such as, for example, category training (Tijsseling and Harnad 1997). This warping effect is a consequence of the restricted number of hidden units and, hence, the necessity of the network to compress redundant information and highlight distinctions with important consequences. An interesting observation is the effect of the identification network on the Garner paradigm tasks. Cross-dimensional interference did not appear when the network performed only categorization, not identification, even in the case where dimensions were integral. An implication of this result is that human subjects in classification tasks may still, perhaps covertly, identify and discriminate stimuli during the acquisition of categories. This is an observation that corresponds with a core principle of exemplar theories (see, e.g. Nosofsky 1992).

We have also suggested a possible mechanism for modelling the established ability of humans to attend selectively to dimensions. The suggested mechanism was derived from Kruschke's ADIT model (1996). The attention-enhanced model was applied to a standard filtration versus condensation task (Kruschke 1991). In a filtration task only one dimension is relevant, whereas in a condensation task both dimensions are relevant. It has been found that subjects find filtration tasks easier to learn because they can attend to one dimension only (Kruschke 1991). We have shown that the attention-enhanced model not only processes filtration tasks more accurately than standard backpropagation, but it also demonstrates an early-block filtration advantage, unlike ALCOVE.

In the Introduction, we discussed that perception contains an initial source of structured information that can be operated upon subsequently by higher cognitive processes Goldstone (1998a), but that these cognitive processes can also modify percepts (e.g. Schyns *et al*. 1998). Several researchers have shown how psychological distances between stimuli further change under the influence of cognitive processes (Goldstone 1994, Goldstone *et al*. 1999). Our simulations with the model are supportive of this bidirectionality: we have shown the richness of perceptual information captured by the Gabor filter component and how this similarity gradient-preserving encoding already captures most of the dimensional interaction. Given these perceptually constrained sensory stimuli, the neural network can develop a psychological space during identification, discrimination and category learning in which representations for stimuli may become warped under task constraints in order correctly to identify, discriminate and categorize the physical stimuli (Tijsseling and Harnad 1997). Given more complex neural networks these learned representations may further develop with additional learning, and we suggest that this may provide a solution to symbol grounding (Harnad 1995, Tijsseling 1998).

The work presented in this paper may have potential relevance to several issues in psychology. For example, by providing a mechanistic explanation of differential processing of dimensions we try to supplement the amount of experimental data collected in the field of cognitive psychology over the last few decennia. This may help point to new directions for experimental research because the accumulated experimental data might simply be insufficient to constrain or guide appropriate theories of dimensional interaction. The simple model of perceptual processing presented here could easily be incorporated into models of other cognitive systems, providing those models with a way to explore interactions between the formation of perceptual representations and other aspects of cognition. For example, in Gluck and Myers' (1993) cortico-hippocampal model, the hippocampal component is modelled with a similar backpropagation network, which also employs predictive differentiation and redundancy compression, argued to be crucial properties of the hippocampal region for learning new information. This model would as such benefit from the incorporation of our model of perception, as this would allow for investigations into phenomena at the intersection of memory and perception, providing predictions for the perceptual and associative processes in category learning.

The model described in the paper supports our claim that connectionist models provide insights into the mechanism of differential processing of dimensions by humans, but it remains a demonstrative model. We adhered to the simplicity principle (Chater 1999) in keeping the model relatively small for purposes of explanation and analysis. The model is also more of a qualitative model in that speed of learning and performance data cannot be quantitatively compared with humans. Studies of simple idealized models are naturally not without drawbacks. One cannot make specific quantitative predictions, as would be expected from more realistic and complex models; but that was not the intent in the first place. Simple idealized models are a means for conceptual clarifications. In this case, the principles and ideology behind the model support our hypothesis that connectionist models can provide a mechanistic approach to dimensional interaction.

There are, however, various directions for future work. For example, we did not apply the model to the standard dimensions of saturation and brightness of a colour (Goldstone 1994), which can be tested with encodings of colour texture using unichrome and opponent features computed from Gabor filter outputs (Jain and

Healey 1997). The effect of increasing stimulus size (e.g. introduce a correlation between a horizontal and a vertical line by widening the lines and therefore increase the number of pixels) and complexity also needs to be researched. We also plan to investigate the role of development in perceptual processing of dimensional interaction. Kovács (2000) showed developmental effects in the visual processing of orientation information across the visual field. Can these developmental characteristics be modelled by systematically adjusting the parameters of the Gabor filter component? For example, manipulating the spatial frequency and reducing the range of the Gabor filter may at some point affect how the interaction between dimensions is captured. Indeed, Kemler and Smith (1978) showed that dimensions that are easily separated by adults are perceived holistically in 4-year-old children. It may be hypothesized that by gradually expanding Gabor filters over images such developmental changes in dimensional interaction can be modelled. To this extent, the properties and structure of Gabor space need to be further systematically explored.

In concluding, we note how Shepard (1964) stressed the importance of finding the transformation that will convert physical interstimulus distances into psychological distances between the corresponding representations. In a similar vein, Potts *et al.* (1998) argue that the traditional distinction between integral and separable should be replaced by models that elucidate the rules that are employed to transform and use multidimensional stimulus variation. These rules are embodied in the sensory processing filter in combination with the processes for redundancy compression and predictive differentiation, and we consequently propose that connectionist models provide this transformation and that analysing their performance will benefit the understanding of human subject performance, concerning which there is abundant behavioural data.

Notes

1. Although 'fused' is a commonly used term to describe integral interaction, it may be deceptive, since—as we shall illustrate later—the psychological dimensions are in fact out of alignment with the physical dimensions and just not reduced to a single dimension.
2. One may suggest injecting random noise in the hidden layer and measuring the resulting output identification error. The probability that adding a specified amount of noise may lead to confusing a representation with another is also a function of the distance between these representations in hidden unit space. However, using noise adds a level of randomness that can make the data less reliable than calculating interstimulus distances directly.
3. Reaction time could also be modelled by training the network for a fixed number of epochs and measuring the residual error. Since the slope of the error curve can vary strongly in steepness for each input, such a method can produce different results for different numbers of epochs. Terminating learning when the error is at the bottom of the slope provides more accurate comparisons.

References

Ashby, F. G., and Maddox, W. T., 1994, A response time theory of separability and integrality in speeded classification. *Journal of Mathematical Psychology*, **38**: 423–466.

Attneave, F., 1963, Dimensions of similarity. *American Journal of Psychology*, **63**: 516–556.

Berlin, B., and Kay, P., 1969, *Basic Color Terms: Their Universality and Evolution* (Berkeley: University of California Press).

Bishop, C. M., 1995, *Neural Networks for Pattern Recognition* (Oxford: Oxford University Press).

Bornstein, M. H., 1987, Perceptual categories in vision and audition. In S. Harnad (ed.) *Categorical Perception: The Groundwork of Cognition* (New York: Cambridge University Press).

Burns, B., and Shepp, B. E., 1988, Dimensional interactions and the structure of psychological space: the representation of hue, saturation, and brightness. *Perception and Psychophysics*, **43**: 494–507.

Buse, R., Liu, Z. Q., and Caelli, T., 1996, Using Gabor filter to measure pattern part features and relations. *Pattern Recognition*, **29**: 615–625.
Cangelosi, A., Greco, A., and Harnad, S., 2000, From robotic toil to symbolic theft: grounding transfer from entry-level to higher-level categories. *Connection Science*, **12** (2): 143–162.
Carroll, J. D., and Arabie, P. 1980. Multidimensional scaling. *Annual Review of Psychology*, **31**: 607–649.
Chater, N., 1999, The search for simplicity: A fundamental cognitive principle? *Quarterly Journal of Experimental Psychology*, **52A**: 273–302.
Daugman, J. D., 1988, Complete discrete 2D Gabor transforms by neural networks for image analysis and compression. *IEEE Transactions Acoustics, Speech, and Signal Processing*, **36**: 1169–1179.
Gabor, D., 1946, Theory of communications. *Journal of the Institute of Electrical Engineering*, **93**: 429–457.
Garner, W. R., 1970, The stimulus in information processing. *American Psychologist*, **25**: 350–358.
Garner, W. R., 1974, *The Processing of Information and Structure* (Hillsdale NJ: Erlbaum).
Garner, W. R., and Felfoldy, G. L., 1970, Integrality of stimulus dimensions in various types of information processing. *Cognitive Psychology*, **1**: 225–241.
Gibson, E. J., 1991., *An Odyssey in Learning and Perception* (Cambridge MA: MIT Press).
Gluck, M. A., and Myers, C. E., 1993, Hippocampal mediation of stimulus representation: a computational theory. *Hippocampus*, **3**: 491–516.
Gluck, M. A., Oliver, L. M., and Myers, C. E., 1996, Late-training amnesic deficits in probabilistic category learning: a neurocomputational analysis. *Learning & Memory*, **3**: 326–340.
Goldstone, R. L., 1994, Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, **123**: 178–200.
Goldstone, R. L., 1995, Mainstream and avant-garde similarity. *Psychologica Belgica*, **35**: 145–165.
Goldstone, R. L., 1998a, Perceptual learning. *Annual Review of Psychology*, **49**: 585–612.
Goldstone, R. L., 1998b, Hanging together: a connectionist model of similarity. In J. Grainger and A. M. Jacobs (eds) *Localist Connectionist Approaches to Human Cognition* (Mahwah NJ: Lawrence Erlbaum Associates), pp. 283–325.
Goldstone, R. L., 1999, Similarity. In R. A. Wilson and F. C. Keil (eds) *MIT Encyclopedia of the Cognitive Sciences* (Cambridge MA: MIT Press), pp. 763–765.
Goldstone, R. L., 2002, Learning to perceive while perceiving to learn. In R. Kimchi, M. Behrmann and C. Olson (eds) *Perceptual Organization in Vision: Behavioral and Neural Perspectives* (Mahwah NJ: Lawrence Erlbaum).
Goldstone, R. L., Lippa, Y., and Shiffrin, R. M., 2001, Altering object representations through category learning. *Cognition*, **78**: 27–43.
Goldstone, R. L., Schyns, P. G., and Medin, D. L. (eds), 1997, *Psychology of Learning and Motivation: Perceptual Learning*, Vol. 36 (San Diego CA: Academic Press).
Goldstone, R. L., Steyvers, M., and Larimer, K., 1996, Categorical perception of novel dimensions. In *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society* (Hillsdale NJ: Lawrence Erlbaum), pp. 243–248.
Goldstone, R. L., Steyvers, M., Spencer-Smith, J., and Kersten, A., 1999, Interactions between perceptual and conceptual learning. In E. Diettrich and A. Markman (eds) *Cognitive Dynamics: Conceptual Change in Humans and Machines* (Cambridge MA: MIT Press).
Gottwald, R. L., and Garner, W. R., 1975, Filtering and condensation tasks with integral and separable dimensions. *Perception & Psychophysics*, **18** (1): 26–28.
Grossberg, S., 1982, *Studies of Mind and Brain: Neural Principles of Learning, Perception, Development, Cognition, and Motor Control* (Boston MA: Reidel Press).
Grossberg, S., 1987, Competitive learning: from interactive activation to adaptive resonance. *Cognitive Science*, **11**: 23–63.
Handel, S., and Imai, S., 1972, The free classification of analyzable and unanalyzable stimuli. *Perception and Psychophysics*, **12**: 108–116.
Hanson, S. J., and Burr, D. J., 1990, What connectionist models learn: toward a theory of representation in connectionist networks. *Behavioral and Brain Sciences*, **13**: 471–518.
Harnad, S., 1987a, Category induction and representation. In S. Harnad (ed.) *Categorical Perception: The Groundwork of Cognition* (New York: Cambridge University Press).
Harnad, S. (ed.), 1987b, *Categorical Perception: The Groundwork of Cognition* (New York: Cambridge University Press).
Harnad, S., 1995, Grounding symbols in sensorimotor categories with neural networks. In *IEE Colloquium 'Grounding Representations: Integration of Sensory Information in Natural Language Processing, Artificial Intelligence and Neural Networks'* (Digest No.1995/103).
Harnad, S., Hanson, S. J., and Lubin, J., 1995, Learned categorical perception in neural nets: implications for symbol grounding. In V. Honavar and L. Uhr (eds) *Symbol Processors and Connectionist Network Models in Artificial Intelligence and Cognitive Modelling: Steps Toward Principled Integration* (New York: Academic Press).
Hock, H., Tromley, C., and Polmann, L., 1988, Perceptual units in the acquisition of visual categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **14**: 75–84.

Hubel, D. H., and Wiesel, T. N., 1962, Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, **160**: 106–154.
Hubel, D. H., and Wiesel, T. N., 1965, Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of Neurophysiology*, **28**: 229–289.
Hyman, R., and Well, A., 1967, Judgments of similarity and spatial models. *Perception and Psychophysics*, **2**: 233–248.
Jain, A., and Healey, G., 1997, Evaluating multiscale opponent color features using Gabor filters. In *Proceedings of the 1997 International Conference on Image Processing* (ICIP '97), Institute of Electrical and Electronics Engineers, Inc.
Johannesson, M., 2001, Combining integral and separable subspaces. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, Edinburgh, 1–4 August.
Johnson, S. C., 1967, Hierarchical clustering schemes. *Psychometrika*, **2**: 241–254.
Jones, J. P., and Palmer, L. A., 1987, An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, **58** (6): 1233–1258.
Kalocsai, P., Zhao, W., and Elagin, E., 1998, Face similarity space as perceived by humans and artificial systems. In *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, pp. 177–180.
Karmiloff-Smith, A., 1992, *Beyond Modularity: A Developmental Perspective on Cognitive Science* (Cambridge MA, Bradford: MIT Press).
Kemler, D. G., and Smith, L. B., 1978, Is there a developmental trend from integrality to separability in perception? *Journal of Experimental Child Psychology*, **26**: 498–507.
Kovács, I., 2000, Human development of perceptual organization. *Vision Research*, **40**: 1301–1310.
Kruschke, J. K., 1991, Dimensional attention learning in models of human categorization. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society* (Hillsdale NJ: Erlbaum).
Kruschke, J. K., 1992, ALCOVE: an exemplar-based connectionist model of category learning. *Psychological Review*, **99**: 22–44.
Kruschke, J. K., 1993, Human category learning: Implications for backpropagation models. *Connection Science*, **5**: 3–36.
Kruschke, J. K., 1996, Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **22**: 3–26.
Laakso, A., and Cottrell, G. W., 1998, How can I know what you think? Assessing representational similarity in neural systems. In *Proceedings of the Twentieth Annual Cognitive Science Conference* (Madison WI, Mahwah: Lawrence Erlbaum).
Lockhead, G. R., 1972, Processing dimensional stimuli: a note. *Psychological Review*, **79**: 410–419.
Mackintosh, N. J., 1975, A theory of attention: variation in the associability of stimuli with reinforcement. *Psychological Review*, **82**: 276–298.
Marcelja, S., 1980, Mathematical description of the response of simple cortical cells. *Journal of The Optical Society of America*, **A70** (11): 1297–1300.
Medin, D. L., and Smith, E. E., 1984, Concepts and concept formation. *Annual Review of Psychology*, **35**: 113–138.
Melara, R. D., Marks, L. E., and Potts, B. C., 1993, Primacy of dimensions in color perception. *Journal of Experimental Psychology: Human Perception & Performance*, **19**: 1082–1104.
Monahan, J. S., and Lockhead, G. R., 1977, Identification of integral stimuli. *Journal of Experimental Psychology: General*, **106**: 94–110.
Murre, J. M. J., Phaf, R. H., and Wolters, G., 1992, CALM: categorizing and learning module. *Neural Networks*, **5**: 55–82.
Myers, C. E., and Gluck, M. A., 1994, Context, conditioning, and hippocampal representation in animal learning. *Behavioral Neuroscience*, **49** (4): 221–227.
Nosofsky, R. M., 1986, Attention, similarity, and the identification-categorisation relationship. *Journal of Experimental Psychology: General*, **115**: 39–57.
Nosofsky, R. M., 1992, Exemplar-based approach to relating categorization, identification, and recognition. In F. G. Ashby (ed.) *Multidimensional Models of Perception and Cognition* (Hillsdale NJ: Lawrence Erlbaum).
Padgett, C., and Cottrell, G. W., 1998, A simple neural network models categorical perception of facial expressions. In *Proceedings of the Twentieth Annual Cognitive Science Conference* (Madison WI, Mahwah: Lawrence Erlbaum).
Pomerantz, J. R., and Garner, W. R., 1973, Stimulus configuration in selective attention tasks. *Perception & Psychophysics*, **14**: 565–569.
Posner, M. I., 1964, Information reduction in the analysis of sequential tasks. *Psychological Review*, **71**: 491–504.
Potts, B. C., Melara, R. D., and Marks, L. E., 1998, Circle size and diameter tilt: a new look at integrality and separability. *Perception & Psychophysics*, **60** (1): 101–112.
Redding, G. M., and Tharp, D. A., 1981, Processing line location and orientation. *Journal of Experimental Psychology: Human Perception and Performance*, **7**: 115–129.
Rumelhart, D. E., Hinton, G. E., and Williams, R. J., 1986, Learning internal representations by error

backpropagation. In D. E. Rumelhart and J. L. McClelland (eds) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (Cambridge MA: MIT Press).

Schyns, P. G., Goldstone, R. L., and Thibaut, J., 1998, Development of features in object concepts. *Behavioral and Brain Sciences*, **21**: 1–54.

Shepard, R. N., 1964, Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, **1**: 54–87.

Shepard, R. N., 1987, Toward a universal law of generalization for psychological science. *Science*, **237**: 1317–1323.

Smith, L. B., and Kemler, D. G., 1978, Levels of experienced dimensionality in children and adults. *Cognitive Psychology*, **10**: 502–532.

Tijsseling, A. G., 1998, Connectionist Models of Categorization: A Dynamical View of Cognition, PhD thesis, Southampton University.

Tijsseling, A. G., and Harnad, S., 1997, Warping similarity space in category learning by backprop nets. In *Proceedings of the Interdisciplinary Workshop on Similarity and Categorization*, University of Edinburgh, pp. 263–269.

Tijsseling, A. G., Pevtzow, R., and Harnad, H., 2002, Dimensional attention effects in humans and neural nets (in preparation).

Torgerson, W. S., 1958, *Theory and Methods of Scaling* (New York: Wiley).

Ward, L. M., 1982, Determinants of attention to local and global features of visual forms. *Journal of Experimental Psychology: Human Perception & Performance*, **8**: 562–581.

Appendix A: Gabor filters

A Gabor function is a complex exponential with a Gaussian modulation. Suppose we have the set of angles $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$, then we can create for each angle α_i a filter F composed of *real* and *imaginary* parts:

$$F_{\Re}(x,y) = e^{-\frac{\left((x-x_0)^2+(y-y_0)^2\right)}{2\sigma^2}} . \sin\left(\tilde{\pi}.\left(\left(\frac{x-x_0}{x_0}\right)\cos\alpha - \left(\frac{y-y_0}{y_0}\right)\sin\alpha\right)\right)$$

$$F_{\Im}(x,y) = e^{-\frac{\left((x-x_0)^2+(y-y_0)^2\right)}{2\sigma^2}} . \cos\left(\tilde{\pi}.\left(\left(\frac{x-x_0}{x_0}\right)\cos\alpha - \left(\frac{y-y_0}{y_0}\right)\sin\alpha\right)\right)$$

in which α= (angle · π)/180°, $\tilde{\pi} = \beta \cdot \pi$ is the spatial frequency of the filter ($\beta > 0$), (x, y) is the current pixel, (x_0,y_0) is the origin or centre of the filter and σ is a constant.

For a given angle, we obtain the response of a particular Gabor filter by calculating the following formula:

$$\sqrt{\sum_{\langle x,y \rangle} F_{\Re}(x,y)^2 + F_{\Im}(x,y)^2}\,.$$

The final input vector to be presented to the network is then the normalized vector of the responses of all Gabor filters for all specified angles.

In simple words, this means that the region surrounding a given pixel in the image is described by the responses of a set of Gabor filters at different frequencies and orientations, all centred at the pixel position. That is, we generate a local description of the image by taking a set of features, which are in fact the responses of a set of Gabor filters distributed over the entire image. This response measures a certain local frequency of the image property at the considered location.

In the simulations we use eight Gabor filters, each tuned to a different angle orientation (0.0°, 22.5°, 45.0°, 67.5°, 90.0°, 112.5°, 135°, 157.5°), centred at and covering the area of the stimulus with a σ between 30.0 and 40.0 in steps of five and a spatial frequency between 2.5 and 3.5 in steps of 0.5 depending on the size of the image, which was exactly as specified in the corresponding papers, providing a total number of filters in the range of eight to 72. The output of all filters was normalized, so that each image would produce a real-valued input vector, the length of which is 1.0 and the number of elements equal to the number of used Gabor filters. The motivation behind these parameters was to restrict the size of the input vectors while at the same time assuring that no information about the visual stimuli is lost. For the explanative purpose of our demonstrative model, we kept the Gabor filter component relatively simple in order to determine how it affects stimulus encoding; although we also ran simulations with a 6 % 6 grid of overlapping Gabor filters for four different

orientations (0°, 45°, 90°, 135°) and varying spatial frequencies, which showed consistent similar results.

The stimuli from Redding and Tharp (1981) had to be encoded with a second set of Gabor filters (one set centred at the left and the other set at the right of the image) to encode variation in the location of the straight line, since using one centred Gabor filter would not capture location differences and produce identical responses for those images varying only in location. To keep the input vector length equal to the other encoded stimulus sets, we used four orientations for this stimulus set.

Appendix B: Mathematical description of attention-enhanced backpropagation

Derivation of attention shifting in backpropagation

As in standard backpropagation, the measure of error is given by

$$E = \frac{1}{2}\sum_i (t_i - o_i)^2 \tag{B 1}$$

in which E is the measure of the error on a particular stimulus, t_i is the target value and o_i is the actual output value of output node i. The total activation arriving at a node j is the sum of all incoming activations o_i multiplied by their attention strengths α_i and the weight on the connection from node i to node j, w_{ij}.

$$net_j = \sum_i w_{ij} o_i \alpha_i \tag{B 2}$$

The activation of node j is then the result of the activation function applied to the net input to node j (the activation function f usually is the sigmoid function, defined as $f(x) = 1/1 + e^{-x}$):

$$o_j = f(net_j). \tag{B 3}$$

We are looking for an algorithm that adapts the attention strengths for each input node and each hidden unit. The attention strength of a node i is adjusted in the opposite direction of the gradient of the error by some constant λ_α, the attention rate parameter:

$$\Delta\alpha_i = -\lambda_\alpha \frac{\partial E}{\partial \alpha_i}. \tag{B 4}$$

To find this gradient we need to use recursive backwards projections from the output layer. We derive this gradient of the error with respect to attention by viewing it as a product of two parts. One part reflects the change in error as a function of the change in net input to the nodes in the next layer and the other part represents the effect that changing attention has on the net input. In other words, to calculate a change in attention of a node j, we use the net input of the nodes to which node j is sending activation to:

$$\frac{\partial E}{\partial \alpha_i} = \sum_j \left(\frac{\partial E}{\partial net_j} \frac{\partial net_j}{\partial \alpha_i} \right). \tag{B 5}$$

There are two cases for nodes *j*, it is either an output node or a hidden unit. So we define:

$$\delta_j = -\frac{\partial E}{\partial net_j}. \tag{B 6}$$

We apply the chain rule on equation (6) to obtain a product of two factors, with one factor reflecting the change in error as a function of the output of a unit and the other one reflecting the change in the output as a function of changes in the input:

$$\delta_j = -\frac{\partial E}{\partial net_j} = -\frac{\partial E}{\partial o_j} \frac{\partial o_j}{\partial net_j}. \tag{B 7}$$

Using equation (B 3) in the second part of equation (7) gives us:

$$\frac{\partial o_j}{\partial net_j} = f'(net_j); \tag{B 8}$$

and using equation (B 1) in the first part of equation (B 7) obtains:

$$\frac{\partial E}{\partial o_j} = -(t_j - o_j); \tag{B 9}$$

from which we further derive:

$$\delta_j = f'(net_j)(t_j - o_j). \tag{B 10}$$

However, when *j* is a hidden unit, we need to use a different derivation as we cannot access the output values directly. The following derivation shows how we can recursively propagate error values back from output to input:

$$\begin{aligned} \frac{\partial E}{\partial o_j} &= \sum_k \left(\frac{\partial E}{\partial net_k} \frac{\partial net_k}{\partial o_j} \right) \\ &= \sum_k \left(\frac{\partial E}{\partial net_k} \frac{\partial}{\partial o_j} \sum_l w_{lk} o_l \alpha_l \right). \\ &= -\sum_k \delta_k w_{jk} \alpha_j \end{aligned} \tag{B 11}$$

Substituting equation (B 11) in equation (B 7) then gives:

$$\delta_j = f'(net_j)\sum_k \delta_k w_{jk}\alpha_j. \tag{B 12}$$

Derivation of the second part of equation (B 5) is as follows. We define the gradient as the effect of a change in attention of a node *i* on the change in net input of all nodes it sends activation to. When node *i* is an input unit, o_i just represents the input value.

$$\begin{aligned}\sum_j \frac{\partial net_j}{\partial \alpha_i} &= \sum_j \frac{\partial net_j}{\partial \alpha_i} \\ &= \sum_j \frac{\partial}{\partial \alpha_i}\sum_l w_{lj} o_l \alpha_l \\ &= \sum_j \sum_l \frac{\partial}{\partial \alpha_i} w_{lj} o_l \alpha_l \\ &= \sum_j w_{ij} o_j\end{aligned} \tag{B 13}$$

Hence, the change in attention strength is calculated according to:

$$\Delta\alpha_i = \lambda_\alpha \sum_j \delta_j w_{ij} o_j. \tag{B 14}$$

Attention rate is dependent on novelty of stimulus
In addition, the attention rate λ_α can be made dependent on the novelty of a stimulus. When a stimulus is new, attention to it should be high enough to enhance learning, but when a stimulus has been presented several times, the novelty wears off and, as such, attention to it will decrease. We define novelty to be related to the amount of error at the output layer. With each presentation of an input the initial attention rate constant is multiplied by the error produced at the output layer:

$$\lambda_\alpha = \lambda_\alpha \left(\sum_i^N \left| t_i - o_i \right| \right), \tag{B 15}$$

in which *N* is the number of output nodes.

Derivation of weight adaptation in attention-enhanced backpropagation
The change of weights is slightly different compared with standard backpropagation in order to incorporate attention strengths. The weight from node *i* to node *j* is adjusted in the opposite direction of the gradient of the error, with the speed of adjustment dependent on the learning rate η:

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}}, \tag{B 16}$$

which can be rewritten as:

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial net_j}\frac{\partial net_j}{\partial w_{ij}}. \tag{B 17}$$

The second factor of equation (B 17) reduces to:

$$\sum_j \frac{\partial net_j}{\partial w_{ij}} = \sum_j \frac{\partial}{\partial w_{ij}} \sum_k w_{kj} o_k \alpha_k = o_i \alpha_i. \tag{B 18}$$

Note that in equation (B 18) attention enters the factor. Now define:

$$\delta_j = -\frac{\partial E}{\partial net_j}. \tag{B 19}$$

We apply the chain rule to obtain a product of two factors, with one factor reflecting the change in error as a function of the output of a unit, and the other one reflecting the change in the output as a function of changes in the input:

$$\delta_j = -\frac{\partial E}{\partial net_j} = -\frac{\partial E}{\partial o_j}\frac{\partial o_j}{\partial net_j}. \tag{B 20}$$

Using equation (B 3) in the second part of equation (B 16):

$$\frac{\partial o_j}{\partial net_j} = f'(net_j). \tag{B 21}$$

Using equation (B 1) in the first part of equation (B 16):

$$\frac{\partial E}{\partial o_j} = -(t_j - o_j); \tag{B 22}$$

from which we derive:

$$\delta_j = f'(net_j)(t_j - o_j). \tag{B 23}$$

When *j* is a hidden unit, we need to use a different derivation as we cannot access the output values directly. Note that since outputs do not have attention strengths, these drop out of the equation (which is similar to assuming that the attention of an output node is equal to 1.0).

$$\frac{\partial E}{\partial o_j} = \sum_k \left(\frac{\partial E}{\partial net_k}\frac{\partial net_k}{\partial o_j} \right)$$

continued . . .

$$= \sum_k \left(\frac{\partial E}{\partial net_k} \frac{\partial}{\partial o_j} \sum_l w_{lk} o_l \alpha_l \right). \qquad \text{(B 24)}$$
$$= -\sum_k \delta_k w_{jk}$$

Substituting equation (B 24) in equation (B 20) then gives:

$$\delta_j = f'(net_j) \sum_k \delta_k w_{jk}. \qquad \text{(B 25)}$$

Hence, the change in weight is calculated according to:

$$\Delta w_{ij} = \eta \delta_j o_i \alpha_i. \qquad \text{(B 26)}$$