Modeling Category Learning
Rodrigues, P.M.A.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
5. On the Neural Correlates of Psychological Spaces: Limits of a View-Based Theory of Object Representation

Current theories of category learning pose the existence of psychological spaces where stimuli are represented as points and point distance determines stimulus dissimilarity. The success of this assumption in explaining human and animal behavior suggests the possibility of a neural counterpart, a brain region or collection of brain regions where neurons encode stimuli in a manner equivalent to the psychological space assumption. Several studies suggest the inferotemporal cortex (IT) to be such a region. In here, we analyze this hypothesis at the light of a view-based theory of visual processing by studying in detail the representation transformations occurring throughout the different processing stages of two models instantiating the theory. We found the hypothesis not to hold in general, by identifying sets of two-dimensional objects for which the representations in IT do not agree with the respective psychological spaces. However, the models considered assume exclusively feedforward processing, whereas the visual cortex is known to have both feedforward and feedback connectivity. Possibly, the hypothesis might hold under feedback processing. Alternatively, psychological spaces might be realized within other brain regions. We discuss both possibilities and suggest a set of computational, behavioral, and neurophysiological studies to elucidate this question.

Similarity plays a fundamental role in current theories of category learning. The more similar two stimuli are, the more difficult it is to assign them to distinct categories. But what makes stimuli similar or dissimilar? In the first attempts to answer this question, research in psychophysics has focused on the relation between perceived similarity and stimulus difference along physical properties, such as the wavelength of light or the frequency of sound (pitch). In most cases, similarity decreased as the physical difference increased. However, the nature of the decrease varied per stimulus set, sensory modality, or animal species. In some cases, similarity even varied in a non-monotonic way. This is the case for tones within one octave (Shepard, 1964). The similarity to the first tone, the one with the lowest pitch, decreases towards the middle of the octave and increases thereafter, forming a U-shaped curve. In short, similarity cannot be unequivocally defined with respect to physical differences. In subsequent explorations, Shepard (1957) proposed the idea of a psychological space where stimuli are represented with respect to a set of dimensions that do not necessarily match the physical dimensions. He pointed out that the psychophysical map realized by the perceptual system does not necessarily translate the physical space in a smooth manner and, therefore, small differences in the physical space might be transformed into large differences in the psychological space and vice-versa. Violations to monotonicity would be confined to the psychophysical map while similarity would vary uniformly with psychological differences. In a seminal paper, Shepard (1987) presented substantial empirical evidence supporting this hypothesis. By assuming merely a monotonic relation between judged similarity and distance within the psychological space, he found that similarity decreased exponentially with distance for a variety of stimulus sets and animal species (humans and pigeons).

Psychological spaces have been adopted by models of identification, recognition, and categorization. The idea of a unique representation space underlying cognitive tasks, has

---

1 This chapter is based on a paper with Jaap Murre as co-author (Rodrigues & Murre, to be submitted)
proved successful in explaining behavior at the psychological level and suggests the possibility of a neural counterpart, a brain region or a set of brain regions that contain the “end product” of the psychophysical map: the psychological space. In here, we consider the case of two-dimensional (2D) visual stimuli, which are commonly used in studies of categorization. Theoretical and neurophysiological results (Cutzu & Edelman, 1998; Edelman, 1995, 1998; Op de Beeck, Wagemans, & Vogels, 2001; Sugihara, Edelman & Tanaka, 1998) suggest that the psychological spaces for visual stimuli are realized within the inferotemporal cortex (IT), the latest stage of visual processing in the cortex. The evidence is, nonetheless, confined to a small number of stimulus sets. In addition, no characterization of stimulus similarity in IT has been proposed yet. In sum, it seems too early to conclude that IT constitutes the neural counterpart of the psychological spaces for visual stimuli in general.

We present a study of stimulus similarity in IT based on neural models instantiating a successful theory of object representation in the visual brain: the view-based theory. We identify stimulus sets whose psychological spaces cannot be found in IT, at least if information is assumed to proceed exclusively in a feedforward fashion, as it is the case in previous theoretical proposals (Edelman, 1995, 1998). However, the visual system is known to possess both feedforward and feedback connectivity. Furthermore, feedback processing is thought to be relevant for “view with scrutiny” (Desimone & Duncan, 1995; Duncan, 1998) by which detailed stimulus representations emerge after visual information has traveled back and forth one or more times within the visual system. Thus, it is possible that the inclusion of feedback might confirm IT as the neural counterpart for psychological spaces. Alternatively, the psychological spaces might be found in brain areas downstream to IT, such as the prefrontal cortex, and result from purely feedforward processing. We discuss briefly these two alternatives suggesting further neurophysiological and behavioral studies.

Psychological spaces and veridicality

A psychological space is an abstraction of the internal representation of objects in the world, which maps perceived stimulus similarity into point proximity in a metric space. Typically, it is derived from similarity judgments (How similar are these two stimuli in a 9-point scale?) or confusion rates in identification tasks, with both empirical measures assumed to vary proportionally to perceived stimulus similarity. The set of all pairwise judged similarities (or confusion rates) is passed to a non-metric multidimensional algorithm (MDS; Kruskal, 1964), which estimates the coordinates of the stimuli in a metric-space by minimizing the mismatch between the rank order of dissimilarity judgments and the rank order of point distances. Spaces of increasing dimensionality are considered in succession until one is found for which the mismatch, quantified in a Stress value, is acceptable and decreases only slightly towards spaces of higher dimensionality. Such space is taken as the psychological space for the stimulus set in question.

Several studies have shown that the psychological spaces of objects varying in shape are veridical (Cortese & Dyre, 1996; Cutzu & Edelman, 1998; Shepard & Cermak, 1973; Sugihara et al., 1998), i.e. the dimensions that define these spaces are a function of intrinsic properties of the objects in the world. For example, in one experiment (Experiment 1, Cortese & Dyre, 1996), human subjects were asked to judge the similarity between every pair of stimuli taken from a set of 9 closed contours. Closed contours can be described by one or more Fourier Descriptors (FDs), each defined by three parameters: frequency, amplitude, and phase. In the experiment in question, the stimuli were generated from three different FDs by
varying the amplitude and frequency of one of them in a combinatorial fashion. The stimuli formed a 3x3 grid in the distal space. The psychological space derived from the similarity judgments resembled closely the distal space with stimuli arranged in an approximately 3x3 grid. In another experiment (Experiment 3, Cortese & Dyre, 1996), the stimuli were again closed contours defined by three FDs, but this time the phase of two of the FDs was varied. The stimuli formed again a 3x3 grid in the distal space. In contrast with the experiment above, the psychological space derived did not resemble the distal space. The stimuli were arranged along one single psychological dimension, which can be defined approximately as the difference between the phases of the two FDs. Thus, although the psychological space differed from the distal space used to define the stimuli, it can still be characterized as a function of the properties of the stimuli and, therefore, veridical. This experiment illustrates the existence of more than one veridical representation for a stimulus set. In fact, every stimulus set accepts an infinitude of veridical representations, as there is an infinite number of mathematical functions that can be applied to one or more stimulus properties. In short, veridicality does not imply predictability. Knowing beforehand which stimulus properties and respective transformations determine the internal representation remains an illusive task (Shepard, 2001).

**Visual processing, veridicality and psychological spaces**

Making sense of visual information is a highly demanding computational task. Part of the difficulty stems from extrinsic factors that affect the appearance of objects, such as the position and pose of the object with respect to the observer or the illumination conditions. To perform object recognition, i.e. to identify an object as previously seen, the visual system needs to be able to ignore the changes induced by such factors when matching the image of the object presented with the traces of the objects stored in memory. Another difficulty comes into play when the visual system has to make sense of a novel object. Similar objects tend to have similar consequences (Quine, 1969) and, therefore, it is often beneficial to respond to a novel object using the response proved effective to a similar object in the past. The difficulty lies on the impossibility of the visual system to access directly the similarity relation between the physical objects. Instead, it has to recover this from the images it receives.

Current theories of visual processing do not yet agree on the way the visual system deals with the difficulties singled out. The structural description theories (Biederman, 1987; Marr & Nishihara, 1978) assume the derivation of a faithful three-dimensional description of the objects. The Recognition By Components theory (RBC; Biederman, 1987) in particular, postulates that every object can be described in terms of a fixed set of primitive volumes called geons and a finite number of spatial relations into which these can enter. Visual processing corresponds to the extraction of the geons and respective arrangement from an object’s image. Because the descriptions are faithful replicas of the original objects, object recognition and object similarity assessment proceed by description matching, a difficult task in itself (Garey & Johnson, 1979). Instantiations of the RBC theory have shown its feasibility with hand-labeled line drawings (Hummel & Biederman, 1992) but, thus far, no model has been able to recover such information from gray-level images. The problem relies on the difficulty of identifying lines and junctions in the image, a fundamental step in the isolation of the geons (Edelman, 1997).

Among the alternative theories, probably the most successful ones are known as view-based theories. Rather than attempting the recovery of a faithful description, these assume
that objects are represented by their similarity to views of objects stored in memory. Recognition follows whenever the similarity to all views of an object stored surpasses a recognition threshold. This scheme has proved successful in computer vision applications (Serre et al., 2007) and could account for psychophysical (Bülthoff & Edelman, 1992) and neurophysiological data (Bülthoff, Edelman, & Tarr, 1995; Logothetis & Pauls, 1995; Op de Beeck et al., 2001) on object recognition. Such a representation scheme has also proved adequate in making sense of novel objects. In a series of articles, Edelman and colleagues have argued and shown that a view-based model of the visual system is capable of veridical representation (Cutzu & Edelman, 1998; Edelman, 1995, 1998; Intrator & Edelman, 1997; Sugihara, Edelman & Tanaka, 1998). The model goes by the name of “Chorus of Prototypes” (Edelman, 1995, 1998) and consists of a (small) set of radial-basis function networks (RBF; Poggio & Edelman, 1990). Each RBF is trained to respond invariantly and maximally to different views of a single object. It consists of a set of units, each selective to a different view and having its response decay with the distance between the current image and its preferred view. The responses of the view-tuned units are combined together to provide the response of the RBF. Intuitively, each RBF computes the resemblance between the view of the object presented and a stored object (prototype). How can such a representation guarantee veridicality? The underlying idea dates back to Shepard (Shepard, 1968) who proposed representation through second-order isomorphism between the relations among external objects and the relations among internal representations. “Thus, although the internal representation for a square need not itself be a square, it should (whatever it is) at least have a closer functional relation to the internal representation for a rectangle than to that, say, for a green flash or the taste of persimmon.” (Shepard & Chipman, 1970, p. 2). Second-order isomorphism presupposes two distinct relation functions, one between real objects in the world (distal similarity) and another one between their internal representations (proximal similarity). Edelman and colleagues assumed that the former could be formalized in terms of a metric space with distal dissimilarity varying proportionally to distance in that space. The latter was assumed to vary with the Euclidean distance between the patterns of response elicited by the Chorus of Prototypes. Veridicality happens whenever small distal distances are mapped to small proximal distances and large distal distances to large proximal distances. According to Edelman and colleagues this is the case for the Chorus of Prototypes.

The mapping from the distal to the proximal space can be understood as a succession of four steps. First, the coordinates of the object in the distal space are translated into its geometry, i.e. its 3D structure. Second, the image of the object (its view) is derived from its 3D structure and from the viewing conditions, i.e. illumination, position and pose of the object relative to the observer. These two steps are extraneous to the model and depend exclusively on the properties of the world. Note that, whereas the first step is likely to preserve the distal distances between the objects, provided the map is smooth, the second might disturb these in a severe way. For example, two human heads might appear identical from the back, because both have a ponytail, but considerably different from the front because one looks Asiatic and the other one European. The relation between objects becomes view dependent. The model adds two extra steps to the mapping. First, the image entering the model is convolved with a number of filters eliciting a representation identical to the one found in V1, the first cortical area in visual processing. Second, the set of RBFs transforms the high-dimensional space representation resulting from the previous step into a small-dimensional one by discounting the effect of pose for those objects represented by the RBFs. Note that not every object in the world is represented by a dedicated RBF. Nevertheless, provided novel objects are sufficiently similar to those stored explicitly, their representation
in the model does not vary substantially with pose. In other words, by learning to respond almost invariantly to different poses of an object, the model will also respond almost invariantly to different poses of objects highly similar to the object learned.

In its current version, the Chorus of Prototypes assumes that all objects are presented at the same retinal position and under the same illumination conditions, with pose being the only factor varying between viewing conditions. As stated above, the model compensates for pose variation by learning invariant representations for some of the objects in the world. The limited number of objects explicitly stored can, however, cause the disruption of the map from distal to proximal similarity. Suppose that only one object was explicitly stored. If two other objects where equally dissimilar to the object stored and dissimilar from each other they would become indistinguishable within the model, an undesirable feat. A possible solution is to store all objects explicitly, but this is for obvious reasons undesirable. The solution adopted by the Chorus of Prototypes, i.e. the storage of only some objects, works fine provided the objects of interest are within the region spanned by the objects stored. The reason has to do with a method called triangulation by which the exact location of a point in a space can be found from its distances to a number of reference points whose coordinates are known in advance. In short, the Chorus of Prototypes seems able to recover a good replica of the distal space for a set of objects provided these are sufficiently similar to those stored and fall within the region spanned by the latter.

Humans and primates are capable of veridical representation and the same holds for the Chorus of Prototypes. A natural question is whether both, animals and model, represent objects in a similar fashion. This question has been addressed by Cutzu and Edelman (1998). In a series of experiments, human subjects were asked to judge pairwise similarity, perform a delayed matching to sample task, and a long-term memory recall task on different classes of computer-rendered 3D animal-like shapes. Each stimulus class formed a characteristic configuration in the distal space (e.g. triangle). A psychological space was derived for each condition and stimulus class. In all cases, the configuration of the stimuli within the psychological space resembled closely the configuration in the distal space. Simulations with the Chorus of Prototypes revealed identical results. The model was trained to respond invariantly to a subset of the objects in a class. Afterwards, the representations elicited by the model for each object in the class were collected and their pairwise distances passed to a non-metric MDS algorithm. The configuration of the objects in the resulting space (the proximal space) revealed significant resemblance with the configuration of the stimuli in the distal space. Thus, both humans and model represent these stimulus classes in identical fashion. A similar study has been carried out with monkeys who had to perform a delayed matching to sample task on a different class of computer 3D animal-like shapes (Sugihara et al., 1998). The configuration of the objects in the psychological space derived from the error rates by non-metric MDS matched closely the configuration built in the distal space. As in the study above, a closely resembling configuration was obtained from the representations of the stimuli in the Chorus of Prototypes through non-metric MDS. In conclusion, at least for the kinds of objects considered in these two studies, the proximal representations derived from the Chorus of Prototypes are in close agreement with the psychological spaces.

Neurophysiology and psychological spaces

The match between psychological- and proximal spaces derived from the representations elicited by the Chorus of Prototypes, suggests a potential mechanism by
which the visual system might solve the problem of object understanding. This poses the existence of cells selective to individual stimuli, responding maximally to a preferred object and having their response decrease with increased distal-space distance between the probe- and preferred objects. A set of such cells selective to different objects should be sufficient to grant a representation equivalent to that of the Chorus of Prototypes and, therefore, guarantee a good match between the neural space derived from the responses of the population of cells through non-metric MDS and the psychological space.

Multiple studies to date have found cells in the inferotemporal cortex (IT), one of the final stages of visual processing in the macaque brain, selective to individual (complex) objects (e.g. Fujita et al., 1992; Gross, Rocha-Miranda, & Bender, 1972; Kotabake & Tanaka, 1994; Tanaka, 1992, 1993), responding maximally to all or a subset of the views of their preferred object and significantly less to other objects (Hasselmo, Rolls, & Baylis, 1989; Logothetis, Pauls, & Poggio, 1995; Rolls, Booth, & Treves, 1996). It has also been shown that both familiarity and explicit training with novel objects induce increased cell selectivity (Freedman et al., 2005; Logothetis, Pauls, & Poggio, 1995; Sakai, Naya, & Miyashita, 1994). These studies, although revealing the existence of stimulus selective cells, do not address the question of whether cell’s response varies proportionally to distal-space distance, a fundamental trait of the Chorus of Prototypes proposal.

In a recent study, Op de Beeck et al. (2001) addressed this question by considering three different groups of 2D closed contours, each containing 8 different stimuli generated from 7 FDs by varying the amplitude of two of them. In all three groups, the stimuli formed a square configuration in the distal space. Neural recordings were taken from cells in the anterior inferotemporal cortex (area TE) of two monkeys while these were watching the stimuli one at a time. In agreement with Chorus of Prototypes, most of the cells revealed a regular tuning for stimuli within the same group, with response decreasing monotonically with distance to their preferred stimulus in the distal space. A neural space was derived for each stimulus group from the population responses. In all three neural spaces, the stimuli were arranged in a configuration resembling the square configuration imposed in the distal space. In another experiment, two other monkeys performed a same-different task on the same stimuli. A psychological space was derived for each stimulus group from the pairwise confusion rates. In all cases, the configuration of the stimuli in the psychological space resembled the square configuration built in the distal space. Thus, a close agreement was found between the neural and the psychological spaces for each stimulus group. Furthermore, this agreement seems to result from a set of cells that mimic the response patterns of object selective units in the model. In short, this study suggests the validity of the Chorus of Prototypes explanation for the emergence of psychological spaces at the level of IT.

The Chorus of Prototypes and its limitations in explaining the emergence of psychological spaces

In all three studies reported above (Cutzu & Edelman, 1998; Sugihara et al. 1998; Op de Beeck et al., 2001), the Chorus of Prototypes has always been able to recover object representations in close correspondence with the psychological spaces. There are, however, reasons to believe that this is not always possible. The problem derives from the transformation taking place at the entry level of the model where the input image is processed by a set of filters resembling simple cells in V1. In the following subsections we present a detailed description of the model, introduce the intuition behind the problem and characterize
some of the conditions in which the transformation at the entry level of the model might interfere with the recovery of the psychological dimensions.

The Chorus of Prototypes: Formal definition

The Chorus of Prototypes is a three-layer feedforward neural network that processes gray-scale images. Units in the first layer resemble simple cells in V1, responding exclusively to the contents of the image falling within a small region denominated the cell’s receptive field and being selective to bars of light at different orientations and sizes. In formal terms, their preferred features are given by their weights, which are defined by a two dimensional Gaussian function elongated along a given orientation,

\[ w_{i,j} = e^{-a(i-i_0)^2 + b(i-i_0)(j-j_0) + c(j-j_0)^2}, \]

\[ a = \left(\frac{\cos\theta}{\sigma_i}\right)^2 + \left(\frac{\sin\theta}{\sigma_j}\right)^2, \quad b = -\frac{\sin 2\theta}{\sigma_i} + \frac{\sin 2\theta}{\sigma_j}, \quad c = \left(\frac{\sin\theta}{\sigma_i}\right)^2 + \left(\frac{\cos\theta}{\sigma_j}\right)^2, \]

where \((i_0,j_0)\) is the position of the center of the receptive field in the image, \((i,j)\) is a position in the image within the receptive field of the unit, \(\sigma_i\) and \(\sigma_j\) are the standard deviation of the Gaussian function along the first and second dimensions, respectively, and \(\theta\) is the angle by which the axes of the function are rotated counterclockwise with respect to the axes of the image. Elongation is achieved whenever the standard deviation values differ from each other. Selectivity is achieved by calculating the normalized inner product between the weights and the gray-scale values of the pixels of the image falling within the receptive field. In formal terms:

\[ y = \frac{\sum_{i,j} w_{i,j} x_{i,j}}{\sqrt{\sum_{i,j} x_{i,j}^2}}, \]

where \(y\) is the response of the unit, \((i,j)\) identifies a position within the receptive field and \(w_{i,j}\) and \(x_{i,j}\) represent the weight of the unit and the content of the image at that position, respectively.

In the simulations reported in here, we considered 160x160 pixel images corresponding to approximately 4.34º of visual angle. The receptive field sizes were varied between 7x7 (0.19º of visual angle) and 39x39 (1.07º of visual angle) in steps of two pixels, in agreement with receptive field sizes reported for parafoveal simple cells in V1 (Schiller, Finlay, & Volman, 1976). We considered four orientations \(\theta=0^\circ, 45^\circ, 90^\circ, \text{ and } 135^\circ\). The values for \(\sigma_i\) and \(\sigma_j\) were defined ad hoc with \(\sigma_j = \sigma_i / 2\) to guarantee elongation and \(\sigma_i = .25\) (receptive field size) + 1.5 to allow the Gaussian function to decay to values close to zero within the receptive field. In total, there are 68 different unit types corresponding to the 17 receptive field sizes and 4 orientations per size. A retinotopic map was obtained for each unit type by having units centered at each pixel of the image sharing the same weights.

Units in the second layer of the network are selective to views of objects. Each unit is connected to all first-layer units and its weights are set to the responses of the first-layer units
to the presentation of the unit’s preferred view. The response is given by a Gaussian function of the distance between the weights and the responses of first-layer units:

\[
y = e^{-\frac{\sum (w_i - x_i)^2}{\sigma^2}},
\]

where \( y \) is the response of the second-layer unit, \( w_i \) is the weight of the connection from first-layer unit \( i \), \( x_i \) is the response of first-layer unit \( i \), and \( \sigma \) is the standard deviation of the Gaussian function.

In the third layer, units are selective to objects. Each unit is connected only to those second-layer units selective to views of the preferred object. The weights are set such that a unit responds almost invariantly and maximally to any view of the preferred object and less for views of other objects. The response is given by the inner product between the weights and the responses of second-layer units:

\[
y = \sum w_i x_i,
\]

where \( y \) is the response of the third-layer unit, \( w_i \) is the weight from the \( i \)-th second-layer unit connected to it and \( x_i \) the respective response. In the cases we consider in here, the objects are two-dimensional and have, therefore, one single view. Consequently, the third layer becomes superfluous and it is ignored in our simulations.

The Chorus of Prototypes translates a gray-scale image into its similarity to a set of objects represented explicitly by units in the third-layer. In all the case studies considered (Cutzu & Edelman, 1998; Sugihara et al., 1998), Edelman and colleagues have shown that there is no need to store every object of a set to guarantee a good recovery of the respective distal space. In general, the objects are arranged within a polygon in the distal space and only those lying at the outermost vertices are assumed to be represented explicitly. This scheme works fine provided the representations of all objects at the first two layers are confined to a “region” surrounded by the representations of those objects selected to be represented explicitly. Preliminary investigations with the two-dimensional stimuli we consider in here have shown that this is not always the case. Thus, we assumed that every object in a set is represented explicitly in the model.

**Recovering a distal dimension**

Before embarking on more complex stimulus sets we will illustrate how the Chorus of Prototypes is able to recover the distal space for some simple two-dimensional stimuli. Let us assume a single vertical line segment centered vertically in the image and whose horizontal position differs across stimuli (see Figure 5.1a). The stimuli are generated from one single distal dimension corresponding to the position of the line segment with respect to the leftmost border of the image. In strict terms, the stimuli do not vary in shape and could, therefore, be judged meaningless for our purposes. However, one could assume the variation in the stimuli to correspond to variation in the relative position of a stimulus part with respect to the remaining parts of a more complex stimulus.
Figure 5.1: Stimulus sets tested on the Chorus of Prototypes. Each row contains three different stimuli belonging to the same stimulus set: a) single vertical line segment varying in the horizontal position in the image; b) smiley varying in the horizontal position in the image; c) two vertical line segments at various distances from each other; d) rectangle with a inner vertical line segment varying in the height of the rectangle and the position of the inner line segment with respect to the left border of the rectangle (only variation of the position of the inner line segment is presented); e) letter E rotated 90° counterclockwise varying in the position of the middle segment from left to right between the other two; f) L-shaped stimuli where both line segments have the same length and a third one departs from the vertex formed by the other two and varies in the angle it forms with the horizontal one between 0° and 90°.
In Figure 5.2 we present the retinotopic map representations obtained at the first layer of the Chorus of Prototypes for receptive field sizes 7, 17, 27, and 37 and orientations 0°, 45°, 90°, and 135° upon the presentation of one of the stimuli. Independently of the map considered, we observe a vertically elongated region of non-zero response located on and around the position of the line segment in the input image. The width of this region increases with the size of the receptive field while the maximum response decreases. Those units selective to vertically oriented bars of light produce the highest responses within each receptive field size. If we would consider the stimulus located immediately to the right of the current one in the distal dimension, we would observe a sliding of the non-zero response region to the right in all retinotopic maps. In fact, the position of the non-zero response regions is in strict correspondence with the position of the stimulus in the distal dimension. In addition, the non-zero response region has its maximum at the center and decays gradually to the sides. This is illustrated in Figure 5.3a where we present the patterns of response produced by the units located along the most central row of the retinotopic map corresponding to receptive field size 39 and preferred orientation 90°. As we show below, these properties are important in the recovery of the distal space from the representations at the first layer of the Chorus of Prototypes.

![Figure 5.2: Responses of units in some of the retinotopic maps at the first layer of the Chorus of Prototypes to the presentation of one vertical line segment. Retinotopic maps sharing the same receptive field size are presented in rows and retinotopic maps sharing the same orientation are presented in columns. Only receptive field sizes 7, 17, 27, and 37 are considered along with orientations 0°, 45°, 90°, and 135°. The whiter the color, the higher the response of the respective unit.](image-url)
In order to derive a proximal space from the representations at the first layer, we need a measure of representation similarity. Intuitively, two representations are similar if all units in the layer produce identical responses. Conversely, the more distinct the responses of individual units are, the more dissimilar the population representations should be. A natural way to encompass this intuition is to define representation dissimilarity as the Euclidean distance between the responses of all units across all retinotopic maps. According to this definition, each individual unit contributes to the overall representation dissimilarity between two stimuli with the square of the difference between the responses produced for each of them. Thus, the more units there are producing different responses, the more dissimilar the representations become. Importantly, not only the number of units producing different responses, but also the magnitude of the response differences are relevant. For example, the representations of stimuli $A$ and $B$ may be more dissimilar than those of stimuli $A$ and $C$ even though there are fewer units producing different responses for $A$ and $B$ than for $A$ and $C$. This is the case if the average squared response difference is larger for stimuli $A$ and $B$ than for stimuli $A$ and $C$.

Let us consider how representation dissimilarity varies with distal-space distance. For the moment, we constrain our analysis to those 160 units located along the most central row of the retinotopic map corresponding to receptive field size 39 and preferred orientation 90°. In Figure 5.3b, we present the response patterns for the first and second stimuli (the two leftmost line segments) elicited on these units. Vertical line segments connect the responses produced by each unit to the two stimuli. Their lengths correspond to the absolute response difference at the single unit level. Note that the absolute response difference varies proportionally to the squared response difference. Thus, the length of a line segment is a good representative of the contribution of the corresponding unit to the overall representation dissimilarity. In Figure 5.3c, we present an identical graph but considering the first and third stimuli instead. We will use these two figures to illustrate why the representations of the first and second stimuli are more similar than the representations of the first and third stimuli.

First, notice that there are fewer units producing a different response for the first and second stimuli, those between positions 47 and 90, than units producing a different response to the first and third stimuli, those between positions 47 and 95. This results from the larger overlap between the non-zero response regions in the former than in the latter case. In addition, the absolute response differences produced by individual units are also smaller in the former than in latter case. To see why, we set units responding positively to at least one of the stimuli in Figure 5.3b in one-to-one correspondence with units in Figure 5.3c. First, let us consider those units that respond more strongly to the first than to the second stimulus. These are found within the dark gray portion of Figure 5.3b. These same units respond more strongly to the first than to the third stimulus (see dark gray portion of Figure 5.3c). In addition, their response to the second stimulus is stronger than that for the third stimulus and, therefore, closer to the response to the first stimulus. Consequently, they present smaller response differences for the first and second stimuli than for the first and third stimuli. Let us now consider those units that respond more strongly to the second than to the first stimulus. These are found within the light gray portion of Figure 5.3b. These units can be set in one-to-one correspondence with units in Figure 5.3c by matching them to units responding as strongly to the third stimulus as they respond to the second stimulus (see light gray portion of Figure 5.3c). Close inspection of the responses of units in the light gray portion of both figures shows that units in Figure 5.3b respond more strongly to the first stimulus than the corresponding units in Figure 5.3c. Thus, also for this subset of units, the response differences for the first and second stimuli are smaller than those for the first and third
stimuli. In conclusion, within the 160 units considered in here, the representations of the first and second stimuli are more similar than the representations of the first and third stimuli. This finding can be extrapolated to all stimuli by noting that stimuli closer together will elicit larger overlap between the non-zero response regions. Thus, there will be fewer units responding positively to at least one of the stimuli than in the case where stimuli are wider apart. Furthermore, for every unit responding differently to the stimuli closer together, there is a unit producing a larger response difference for the stimuli wider apart. In sum, larger distances between the stimuli in the distal space induce larger representation dissimilarities.

Figure 5.3: Responses of units disposed along the most central row of the retinotopic map corresponding to receptive field size 39 and orientation 90º in the Chorus of Prototypes to stimuli in the vertical line segment stimulus set: a) responses to all stimuli; b) responses to the first and second stimuli; c) responses to the first and third stimuli. In both b) and c) panels, the response difference produced by each unit is highlighted by a vertical line connecting both responses. The shaded areas in these two panels are meant to help in understanding why the overall response differences are larger between the first and third stimuli than between the first and second stimuli. See text for more details.
Up to now, we considered the representations elicited within only 160 units of the 1,740,800 (=17 receptive field sizes x 4 orientations x 160 x 160) units found in the first layer of the Chorus of Prototypes. However, the same analysis can be applied to every row of units in every retinotopic map as, in all cases, the presentation of a stimulus elicits an identical Gaussian response pattern centered at the position where the stimulus is located. Thus, representation similarity varies proportionally to distal space distance even when all units in the first layer of the model are considered. Recall that the non-metric MDS algorithm relies on the rank order between similarities to establish the coordinates of the stimuli in a metric space. As we have shown, the rank order of representation similarities is the same as the rank order between the distal space distances. Thus, it is not surprising that the space derived from the representation similarities (proximal space) is identical to the distal space.

Before analyzing the proximal space derived from representations at the second layer of the model, a note of caution on the relevance of receptive field sizes on the recovery of the distal space is worthwhile. The receptive field sizes vary substantially and the largest is still smaller than the size of the input image. As we noticed above, the larger the receptive field size, the wider the non-zero response regions. In our detailed analysis, we considered units having the largest receptive field size. The overlap between the non-zero response regions generated by these units decreases strictly with distal distance between the stimuli, inducing a strict decrease of representation similarity with distal space distance. This is not the case for smaller receptive field sizes, for which the non-zero response regions are narrower, causing overlap to decrease up to a given distal distance and become non-existent for larger distances. In this case, dissimilarity still varies monotonically with distal distance but not strictly monotonically. Consequently, the space derived through non-metric MDS from these dissimilarities only does not resemble the distal space. Nevertheless, when considering all retinotopic maps simultaneously, the space derived does resemble the distal space. Thus, if the largest receptive field were narrower than the width of the portion of the input image spanned by the stimuli, the space recovered might differ from the distal space.

At the second layer of the model, there are as many units as there are stimuli in the stimulus set, each selective to a different stimulus. The patterns of response for each stimulus are presented in Figure 5.4. Contrary to the case of the first-layer, the representations are not simply transposed versions of each other. Those stimuli close to the extremes of the distal dimension are represented by a truncated version of the response pattern associated with the stimulus in the middle. This brings important consequences to representation dissimilarity as measured by Euclidean distance. The representation of the stimulus at one of the extremes of the distal dimension is more similar to the representation of the stimulus at the opposite extreme than to the representation of the stimulus just before the opposite extreme. This is reflected in the one-dimensional space derived through non-metric MDS where the first stimulus appears between the second and the third and the last between the 5th and 6th. The remaining stimuli are equally spaced and appear according to distal order. The problem is that stimuli closer to the extremes of the distal dimension are represented by vectors of responses with a smaller length than those of stimuli at intermediate positions. This problem is circumvented by defining stimulus dissimilarity as the Euclidean distance between the normalized responses. The non-metric MDS derived using this definition recovers a perfect replica of the distal space. Because this problem is not specific to the stimulus set considered, we recur to this same definition of stimulus dissimilarity whenever considering representations at the second layer. In conclusion, the Chorus of Prototypes is able to recover a perfect replica of the distal space for the single line segment stimuli.
Chapter 5

Figure 5.4: Responses of units in the second layer of the Chorus of Prototypes to all stimuli in the single line segment stimulus set.

The stimulus set just considered matches approximately the preferred feature of some of the units in the first layer of the Chorus of Prototypes. More specifically, that of units having their weights defined by a Gaussian function elongated vertically. One might argue that this is the reason why the distal space could be recovered by the model. However, as we have shown above, even those units having other preferred features contribute positively to the recovery of the distal space. This suggests that the model might be able to recover the distal space for more complex stimuli varying in position in the visual field. We tested this hypothesis by considering a set of 7 smileys defined within a 21x21 pixel patch and varying in their horizontal position in the visual field by steps of 5 pixels (see Figure 5.1b).

Figure 5.5: Responses of units in four of the retinotopic maps in the first layer of the Chorus of Prototypes to the presentation of smiley stimuli. The retinotopic maps share the same preferred orientation (90°) and are arranged from left to right by increasing receptive field size (7, 17, 27, and 37 pixels). Higher responses are represented by whiter colors. a) responses to the presentation of a single smiley; b) squared response differences for the first and third stimuli; c) squared response differences for the first and fifth stimuli.
In Figure 5.5a, we present the responses elicited at the retinotopic maps corresponding to receptive field sizes 7, 17, 27, and 37 and orientation 90° for the first stimulus. The smiley appears with decreasing detail from the smallest to the largest receptive field size. The maximum responses are obtained for the smallest receptive field size and decrease towards the largest one. As the smiley slides from left to right in the image so does the corresponding representation in all retinotopic maps. This induces increasing differences in the responses due to the decreasing overlap between the regions of non-zero response (see Figures 5.5b and 5.5c). The end result is a monotonic relation between the dissimilarities measured at the first layer of the model and the distal distances. The space derived through non-metric MDS from these dissimilarities is a perfect replica of the distal space. The same holds for the space derived from the representation dissimilarities at the second layer.

The results in this section suggest that the Chorus of Prototypes is able to recover the distal space if this corresponds to variation in position of a whole stimulus in the visual field. In the next section, we consider a more complex case where variation in position occurs for a stimulus part with respect to other stimulus parts.

**Recovering a distal dimension under stimulus part interference**

Units in the first layer of the Chorus of Prototypes match whatever falls within their receptive fields with their preferred feature. A key aspect in the matching process is the normalization by the amount of activity entering the unit (see Equation 2), which allows it to be selective to the same feature irrespectively of the strength of the incoming activity. As we shall see below, this also denies the Chorus of Prototypes the ability to recover a good replica of the distal space under certain conditions. In the present case, we consider a set of 7 stimuli made out of two vertical line segments with the same length, aligned vertically and at a given horizontal distance from each other (see Figure 5.1c). The stimuli are defined in terms of one single distal dimension corresponding to the distance between the line segments. In our simulations, we assumed the stimuli to appear in the input image aligned by the leftmost line segment.

In Figure 5.6a we present the matrix of pairwise dissimilarities derived from the responses of all units in the first-layer of the model along with a matrix of pairwise dissimilarities that vary monotonically with the distal dimension, both presented in a graph format. The two matrices differ substantially. In the former, dissimilarity does not always increase with distal distance. As an example, the dissimilarity to the last stimulus (stimulus 7) increases from the first to the fourth stimulus and decreases thereafter. Similarly, the dissimilarity to stimuli 4 through 6 increases from the first stimulus up to a stimulus in between and decreases thereafter. We derived a space through non-metric MDS on the dissimilarity values (see Figure 5.6b). As expected, the stimuli are not arranged along a single dimension. Instead, they form a spiral with the first stimulus at the origin and the remaining stimuli disposed along the spiral according to distal order.

To understand why representation dissimilarity does not vary proportionally to distal distance, we pass to a detailed analysis of the nature of the representations. On every retinotopic map at the first layer of the model, the stimuli are represented by one or two vertically elongated regions of non-zero response located on and around the positions of the line segments in the image. A single region is obtained whenever the size of the receptive field is large enough to encompass both lines in the stimulus and these are sufficiently close together. In Figure 5.7 we present the responses of those units with receptive field size 17 and
orientation 90° located along the central row of the respective retinotopic map. All stimuli induce two response peaks but the ones for the first stimulus are much lower than the remaining ones. This is a consequence of normalization at the single unit level. The receptive fields of those units centered on one of the line segments of the first stimulus cover not only a portion of the line segment on which they are centered but also a portion of the other line segment. In contrast, the receptive fields of the units centered on one of the line segments of one of the remaining stimuli do not cover any portion of the other line segment in the stimulus. The normalization value is thus higher for those units responding to the line segments in the first stimulus than for the corresponding units in the remaining stimuli. This has important consequences to stimulus dissimilarity. The response patterns for the first and last stimuli differ at those units covering the leftmost line segment in both stimuli, those covering the rightmost line segment in the first stimulus and those covering the rightmost line segment in the last stimulus. The differences produced by the first two sets of units are small in comparison to the differences produced by the last set of units (see Figure 5.7). In the case of the second and last stimuli, only those units covering the rightmost line segment in the second stimulus and those covering the rightmost line segment in the last stimulus produce different responses.

Figure 5.6: a) Graphical representation of the stimulus dissimilarity matrix obtained for the two vertical lines stimulus set from the responses of units in the first layer of the Chorus of Prototypes with normalization (left) and without normalization (right). b) spaces derived through non-metric MDS from the dissimilarities presented in panel a).
There are thus fewer units producing different responses for the second and last stimuli than for the first and last stimuli. However, the overall differences for the second and last stimuli are larger than those for the first and last stimuli. The end result is a larger similarity between the first and last stimuli than between the second and last stimulus. This result holds for the full retinotopic map as the remaining rows of units produce identical representations. In the case of larger receptive field sizes, not only the first stimulus, but also those immediately following have their peak responses reduced by higher normalization values. Assume this is the case for the first two stimuli. In an analysis analogous to the one just presented, one concludes that the second stimulus becomes more similar to the last stimulus than the third stimulus. Thus, violations to monotonicity between stimulus dissimilarity and distal distance are present across retinotopic maps. To confirm that normalization and not the preferred features of first-layer units are the cause of the problem, we simulated the model with normalization turned off. The matrix of pairwise dissimilarities obtained from unit responses is presented in Figure 5.6 along with the space derived through non-metric MDS. The distal space is recovered perfectly.

At the second layer, there are only 7 units, one per stimulus. As in the previous section, the space derived through non-metric MDS on the dissimilarities between second-layer representations is a perfect replica of the space derived from first-layer dissimilarities. In other words, the second layer retained the relations between the stimuli induced at the first layer although it represents the stimuli in a much lower dimensional space.

This case study shows that the Chorus of Prototypes has problems in recovering the distal dimension when this translates into variation in the relative position of one stimulus part with respect to another. The analyses presented in this and in the previous section suggest that this might not always be the case. We have seen how the position of a stimulus can be recovered provided the largest receptive fields cover the portion of the visual field spanned by the stimuli. In the case of the two line stimuli, if the leftmost line segment (the fixed part) would be set further to the left it would be possible to have units whose receptive fields cover the full range of variation of the right line (the moving stimulus part) without covering the left one. Note that there could still be units whose receptive fields cover both lines in the stimuli and, therefore, induce effects such as those presented above. Nevertheless, we expected these to be less pronounced. A larger receptive field can cover more stimuli close to its borders where the weights, which are defined by a Gaussian function, are still close to zero.

Figure 5.7: Responses of units disposed along the most central row of the retinotopic map in the Chorus of Prototypes corresponding to receptive field size 17 and orientation 90° to all stimuli in the two vertical lines stimulus set.
The effects of normalization are, therefore, less pronounced or even absent. If present, they might even be compensated by the dissimilarities induced by the smaller receptive fields. We tested this hypothesis in a new stimulus set. The stimuli are again formed by two vertical line segments with the same length, aligned vertically and at increasing distances from each other. To guarantee that there would still be receptive fields covering both lines in all stimuli, we have chosen for distances 14 through 26 in steps of two pixels. The overall dissimilarities measured at the first-layer are in monotonic relation with the distal distances. Only in 3 of the 68 retinotopic maps did we find a violation of monotonicity with the first stimulus being more similar to the last one than the second stimulus. The space derived through non-metric MDS from these dissimilarities resembles the distal space. The only difference resides in the increasing spacing between consecutive stimuli from the first to the last pair. The space derived from the dissimilarities in the second layer is also one-dimensional and the stimuli appear ordered according to the distal order. Their spacing, however, increases from the extremes towards the center. This change in spacing is probably solved by having units in the second layer represent stimuli located at positions in the distal dimension beyond those of the most extreme stimuli in the stimulus set. In any case, the model performs reasonably well in recovering the distal space.

**Problems in recovering psychological dimensions**

In the previous section, we have shown how the Chorus of Prototypes has problems in recovering a distal dimension of variation if this corresponds to the relative position of a stimulus part with respect to another stimulus part and if the two are too close to each other. This is the case in the “line in rectangle” stimulus set commonly used in studies of category learning (e.g. Erickson & Kruschke, 1998, 2002b; Kruschke, 1993). The stimuli consist of a rectangle with a vertical line segment inside located close to the base. They vary along two dimensions: the height of the rectangle and the horizontal position of the inner line, which spans the full width of the rectangle. The psychological space derived by non-metric MDS from the similarity ratings collected from human subjects is two dimensional, with the stimuli arranged in a grid fashion just as in the distal space (Erickson & Kruschke, 1998; Kruschke, 1993).

We simulated the Chorus of Prototypes with this stimulus set. The rectangles were aligned by their lower-left vertex. They are 41 pixels wide and vary in height between 11 and 41 pixels in steps of 5. The line segment is 7 pixels tall and its position, as measured from the leftmost border of the rectangle, varies between 5 and 35 pixels in steps of 5. In total, there are 49 stimuli corresponding to all combinations of 7 heights and 7 line positions. The dissimilarities between the representations at the first layer are more pronounced for variations along the height dimension (between 0 and 57.6) than for variations along the line position dimension (between 0 and 14.1). This results from the fact that variation along the height dimension induces response changes in many more units than variation along the line position dimension as the former consists of a vertical transposition of a line segment 41 pixels wide and the latter consists of the horizontal transposition of a line segment 9 pixels tall. When considered in isolation, the dissimilarities between stimuli varying along the height dimension vary monotonically with height difference. Monotonicity is not obtained in the case of variation along the line position dimension. Here, the dissimilarity increases up to a point and decreases thereafter. As a result, the stimuli whose line segment is just next to the
left and right borders of the rectangle are more similar to each other than to the stimulus having its line segment at the most central position.

The monotonicity found for the case of height variation is surprising in the light of the results of the previous section. The distance between the top and bottom borders of the smallest rectangle is not larger than the distance between the top borders of the smallest and tallest rectangles, a condition we argued to be necessary to guarantee monotonicity. Here, however, the left and right borders of the rectangle increase with increased height. This guarantees that there are more units responding to a taller rectangle than to a smaller one, which compensates for the violations to monotonicity produced by those units whose receptive fields cover the bottom and top borders simultaneously. In contrast, the lack of monotonicity in the case of line position variation agrees with the analyzes in the previous section. Those stimuli whose line segment is close to the borders have the representation of the line suffer more from interference, due to higher normalization values, than those with the line closer to the center.

All the properties of the dissimilarity values here summarized are reflected in the space derived from non-metric MDS. The space is one-dimensional and the stimuli are arranged by height with stimuli having the same height present at nearby positions. The same space is obtained when considering the dissimilarities taken from second-layer representations. We also derived a space from the dissimilarities measured at the first and second layers for each subset of stimuli varying in line position. With the exception of those subsets associated with the two smallest heights, a two dimensional space produced a fair representation for the stimuli, with the stress value smaller than .081. The stimuli form a closed contour with the first and last stimuli next to each other and the remaining ones ordered according to line position along the contour (see Figure 5.8).

Figure 5.8: Spaces derived through non-metric MDS from the dissimilarities of the line in rectangle stimuli measured at the first layer of the Chorus of Prototypes. One space was derived for each subset of stimuli sharing the same rectangle height. The spaces are organized by increasing rectangle height from left to right and top to bottom.
These results show that the model is unable to recover a good replica of the psychological space for this particular stimulus set. It does, however, work well with other stimulus sets commonly used in studies of category learning such as rectangles varying in height and width (Maddox & Ashby, 1993), semicircles varying in radius containing a radial line varying in angle (Ashby & Lee, 1992; Nosofsky, 1985, 1989) or inverted L shapes varying in the length of each line segment (Ashby & Gott, 1988; Maddox, Ashby, & Waldron, 2002). Nonetheless, this does not mean that there are no other important stimulus sets for which the model is unable to recover a good replica of the psychological space. We have devised two stimulus sets for which no study has yet estimated the nature of the respective psychological spaces. The first consists of a set of E letters rotated 90° counterclockwise where the position of the middle line segment varies from left to right, spanning the full space between the other two vertical line segments (see Figure 5.1e). The second consists of a set L-shaped stimuli where both line segments have the same length and a third one departs from the vertex formed by the other two and varies in the angle it forms with the horizontal one between 0° and 90° (see Figure 5.1f). In both cases, we expect the psychological space to be one-dimensional with stimuli arranged according to the relative position of the middle line segment with respect to one (and only one) of the extreme line segments, just as in the case of the line in rectangle stimuli. We simulated the model with both stimulus sets and, in both cases, we obtained a two dimensional space through non-metric MDS on the dissimilarities measured at the first and second layers of the model. The stimuli formed a curved contour with the first and last stimuli close together and the remaining ones arranged along the contour according to distal order (see Figure 5.9). This is the result of violations to monotonicity between the dissimilarities among stimulus representations and the expected psychological distances, which are induced by interference between the middle line segment and the most extreme ones whenever the former gets too close to the latter. If our hypothesis holds, the Chorus of Prototypes has also problems in recovering the psychological spaces for these stimulus sets.

**Figure 5.9:** Spaces derived through non-metric MDS from the dissimilarities measured at the second layer of the Chorus of Prototypes for the counterclockwise rotated E’s (left) and the L-shaped stimulus set containing a middle line segment varying in angle with respect to the horizontal line segment (right).

### Beyond the Chorus of Prototypes

The Chorus of Prototypes has been proposed as a proof of concept model for the idea of representation through second-order isomorphism. It collapses the hierarchy of processing stages known to take place in the visual brain by assuming only two stages, the first corresponding to the action of simple cells in V1 and the second corresponding to the action
of cells in IT selective to complex objects. Intermediate stages in the visual brain are thought relevant in obtaining position and scale invariance, a feat the Chorus of Prototypes lacks. The effects produced by the inclusion of such intermediate stages in the recovery of psychological spaces remains an open question. They might induce a poor recovery of the psychological space in cases in which the Chorus of Prototypes performs well. Alternatively, they might be of help in cases where the Chorus of Prototypes has difficulties. In the next section, we address these questions by considering a recent model of shape processing in the visual brain: the Standard Model (Riesenhuber & Poggio, 1999; Serre, Wolf, & Poggio, 2004; Serre et al., 2005).

The Standard Model and the recovery of psychological spaces

The Standard Model

The Standard Model accommodates, arguably, the main physiological, anatomical and biological constraints of the feedforward path of the “what” stream in the visual brain, which we will summarize briefly. Visual information enters the mammalian visual system through approximately 120 million photoreceptors arranged in a two-layer sheet on the back of the eye that translate reflected light into electrical activity. The optical nerve transports action potentials from ganglion cells in the eye to the lateral geniculate nucleus (LGN), the first brain area receiving visual information. Cells in the LGN have a center-surround structure, responding to differences of light intensity. In the next stage, the primary visual cortex (V1), simple cells respond primarily to light gratings with a specific spatial frequency and orientation presented at a particular position within the receptive field. Complex cells are also primarily responsive to light gratings but their response does not depend on the exact location of the preferred stimulus within the receptive field. Processing of visual information proceeds thereafter along two pathways, one mostly dedicated to shape processing, the “what” stream, and another one mostly concerned with the processing of movement and location, the “where” stream. Here, we are especially interested in the “what” stream. This departs from V1 and traverses areas V2 and V4 until reaching IT. At each subsequent area, cells have larger receptive fields and are selective to increasingly complex shapes. In IT, for example, cells have receptive fields covering the full visual field and are selective to faces (Desimone et al., 1984; Gross et al., 1972; Hasselmo et al., 1989; Perret et al., 1992) and other complex objects (Logothetis et al., 1995; Sakai et al., 1994). Their responses are only slightly affected by translation and scale transformations of their preferred stimulus (Logothetis & Sheinberg, 1996; Logothetis et al., 1995; Perret & Oram, 1993; Tanaka, 1996). They are also relatively, and sometimes completely, insensitive to rotation in depth, responding invariantly to some or all views of their preferred stimulus. View invariance is thought to rely on stimulus specific experience. In a compelling experiment, Logothetis et al. (1995) trained monkeys to recognize a set of isolated views of 3D wireframe structures. The stimuli were arguably new to the monkeys. When testing the response of IT cells, Logothetis et al. found that the great majority of cells selective to the training objects were view-tuned, with invariance in depth-rotation constrained to 20 degrees from the trained view in both directions.

Every two consecutive areas in the “what” stream have feedforward as well as feedback connectivity. It is widely accepted that the first 150 ms of visual processing rely mostly on information communicated in a feedforward manner. Stimulus dependent activity arrives cells in IT 80-100ms after stimulus onset (Keysers et al., 2001; Perret et al., 1992), a
too narrow time window to allow information to be backpropagated to lower areas and return back. Nevertheless, there is sufficient and robust information in the firing rate of IT cells just after 100ms of stimulus onset to perform a number of recognition tasks (Hung et al., 2005). This is corroborated by behavioral studies that show that monkeys and humans are able to perform recognition and categorization judgments under rapid presentation of visual scenes (Potter, 1975; Thorpe, Fize, & Marlot, 1996).

The Standard Model extends the idea of simple and complex cells found by Hubel and Wiesel (Hubel & Wiesel, 1962, 1965) in V1 to the remaining processing stages in the “what” stream, by interleaving simple cell- with complex cell stages a number of times. Simple units (S units) in the model provide selectivity and generalization by responding maximally to a preferred stimulus and progressively less for stimuli at increasing dissimilarity. Their response is defined just as the response of second layer units in the Chorus of Prototypes by a Gaussian function of the Euclidean distance between the weights, which establish the preferred feature, and the incoming pre-synaptic activity (see Equation 3). Complex units (C units) provide transposition and scale invariance mechanisms. They pool over a set of S units preferring the same feature at slightly different locations and slightly different scales. Their response is the maximum over the responses of all S units in the pool.

In the most recent version of the model (Serre et al., 2005), two dissociated routes within the “what” stream are distinguished. One corresponding to the full path between V1 and IT, named the main route, and one corresponding to the V1->V2 ->IT path, bypassing V4: the bypass route. We performed some preliminary simulations with the full model on the stimulus sets considered above and found that the largest representation differences in IT were produced via the bypass route. This means that the bypass route dominates representation dissimilarity in IT. Arguably, the main route exerts a minimal influence and can, therefore, be ignored. We tested this hypothesis by applying a non-metric MDS to the representations in IT obtained within the full model and within the model constrained to the bypass route. The stimulus configurations obtained in both cases were barely distinguishable, confirming the minimal influence of the main route. Therefore, we decided to constrain our analysis to the bypass route. From here onwards we will use Standard Model to mean the model constrained to the bypass route.

We pass to a short description of the model. For more details and parameter values, please consult Serre et al. (2005, p. 69-73). The model contains five different stages, S1, C1, S2b, C2b, and S4. The first letter of a stage stands for the type of units found in it. A tentative mapping of model’s layers into brain areas presented in Serre et al. (2005) sets S1 units and simple cells in V1 in correspondence. C1 units represent complex cells found in V1 and possibly V2. S2b and C2b units represent cells found in posterior and anterior IT. Finally, S4 units correspond to view tuned cells found in anterior IT.

The input to the model is a gray scale image of 160x160 pixels, corresponding to approximately 4.34º of visual angle. The activity at the S1 stage is obtained by convolving the image with a set of Gabor filters varying in orientation, spatial frequency, and size. As in the case of the Chorus of Prototypes, each S1 unit normalizes the incoming activity before computing the inner product with its weights to guarantee the preservation of selectivity under different light conditions. For simplicity reasons, only 17 sizes, from 7x7 (0.19º of visual angle) to 39x39 (1.07º of visual angle) pixels in steps of two pixels and four orientations (0º, 45º, 90º, and 135º) are considered, with spatial frequency varying inversely proportional to filter size. In total, S1 contains 68 (=17 sizes x 4 orientations) different representations of the input image, each a retinotopic map of 160x160 units. These maps are
arranged into 17 bands, one per receptive field size, each containing the 4 retinotopic maps corresponding to the 4 possible filter orientations. Intuitively, each band represents a set of features (oriented gratings of light) at a specific scale and each feature is represented across bands at different scales. For the purpose of exposition, we assume the bands to be ordered by filter size.

The 17 bands in S1 are transformed into 8 bands in C1 by bringing together pairs of consecutive bands, with the exception of the last three bands that are all brought together. To illustrate how this is realized, we consider the case of the first band in C1, which brings together bands 1 and 2 in S1. The band contains 4 retinotopic maps, one for each filter orientation. Every single unit in a retinotopic map receives input from all units in the corresponding retinotopic maps in bands 1 and 2 that are localized within a neighborhood of its position. For example, the unit at position (80,80) in the retinotopic map associated with filter orientation 0º receives input from all units localized within a neighborhood of (80,80) in the retinotopic map associated with filter orientation 0º in band 1 and all units falling within the same neighborhood in the retinotopic map associated with the same filter orientation in band 2. This connectivity scheme is used across all C1 bands. In conjunction with the MAX operation, the connectivity scheme makes C1 units selective to oriented gratings and, simultaneously, slightly invariant to position and scale variation. Their receptive field sizes become also larger than those of S1 units as they respond to anything falling within and between the receptive fields of all incoming S1 units. For efficiency reasons, the retinotopic maps in C1 are downsampled by representing only those units at regular intervals from each other, with the width of the interval increasing with band number.

In the S2b stage, units become selective to more complex shapes. Selectivity develops by passive exposition to a set of (natural) images. In an initial training phase, an image is presented and the responses of C1 units are computed. One band and one location within the respective retinotopic maps are selected randomly. The responses of some of the C1 units falling within a $N\times N$ grid, with $N$ denoting the width of the grid in a retinotopic map, centered at the location chosen and spanning across the 4 maps in the band, become the weights of a new S2b unit-type. This unit-type becomes, thus, tuned to one feature in the training image. It will respond maximally whenever that feature is present in the current image and less to other features. The process is repeated a number of times until a sufficient number of S2b unit-types is obtained. Just as the C1 stage, the S2b stage is also organized into 8 bands, each corresponding to a different scale. In each band, there are as many retinotopic maps as the number of S2b unit-types. Thus, whereas C1 represents only 4 different features at different scales, S2b represents a multitude of complex features at those same scales. All units in a retinotopic map share the same weights, those of the corresponding S2b unit-type. The response of a S2b unit is computed with respect to the responses of the $N\times N$ grid of C1 units centered at the position of the S2b unit and spanning the 4 retinotopic maps in the band.

In the C2b layer, there are as many units as the number of S2b unit-types. The activity of each unit is given by the maximum of the activities of all S2b units of the same unit-type, independently of location (within a retinotopic map) or scale (band). C2b units are, therefore, position invariant and considerably scale invariant.

In the final stage of the model (S4), units are selective to particular views of the objects. In contrast with the S2b stage, learning occurs in a supervised manner. For each view, a new unit is added and its weights are set to the response of (some of) the C2b units to the image of that particular view. As in the Chorus of Prototypes, view invariance is obtained by linear combination of a set of S4 units selective to different views of the same object.
Recovering psychological dimensions

We simulated the Standard Model with a subset of the line in rectangle stimuli. All stimuli have constant rectangle height (36 pixels) and vary only on the position of the inner line segment. The features associated with the S2b unit types were extracted from the training stimuli themselves and not from natural images, as it is commonly the case. The reason is purely computational. Units selective to features barely resembling the ones present in the stimuli produce small response differences across stimuli. The opposite takes place for those units selective to resembling features. The differences between stimulus representations, as measured at the population level, are thus almost exclusively dependent on the latter units. We can, therefore, decrease the number of S2b unit types without distorting the differences obtained. In total, we extracted 240 different features. In S4, we assumed as many units as stimuli in the stimulus set, each selective to a different stimulus.

In Figure 5.10, we present the space derived through non-metric MDS on the stimulus dissimilarities measured at S4. As in the case of the Chorus of Prototypes, we assumed dissimilarity to be defined by the Euclidean distance between the normalized representations at the latest stage of the model. The space is two-dimensional with the stimuli forming a curved contour when traversed by stimulus distal order. The stimuli in the first half of the distal dimension are at progressively larger distance from the first one. Once the middle stimulus is passed, the stimuli start approaching the first stimulus. Thus, the model does not recover a good replica of the psychological space. Below we analyze why this is the case.

The representations in S1 are slightly different from those in the first layer of the Chorus of Prototypes. Let us consider the representation elicited by a single vertical line segment in the retinotopic maps containing units preferring vertically elongated gratings of light (see Figure 5.11). The non-zero response regions are elongated vertically and centered at the position of the vertical bar in the input image, just as in the Chorus of Prototypes. However, whereas in the Chorus of Prototypes the non-zero response regions peak at the center and decay gradually towards the sides (see Figure 5.2), in the Standard Model they alternate between increase and decrease phases from side to side, with the peak values increasing from side to center. This “wavy” pattern is a consequence of the Gabor function defining the weights of S1 units, where positive and negative value regions are interleaved orthogonally to the preferred orientation. The existence of multiple peaks and valleys within the non-zero response regions can, in certain conditions, disrupt the ideally monotonic relation between representation dissimilarity and distal distance. To illustrate this, we make use of the vertical line segment stimulus set introduced before, where stimuli are single vertical line segments varying in their horizontal position. In Figure 5.12, we present the response patterns elicited by the first three stimuli, those located at the three leftmost positions in the image, on those units spread along the most central row of a retinotopic map selective to vertically elongated gratings of light. In the top panel, the response patterns elicited by the first and second stimuli are presented. Notice how the peaks and valleys of the response pattern corresponding to the first stimulus are aligned with the valleys and peaks, respectively, of the response pattern corresponding to the second stimulus. This results in large response differences within the portion were both non-zero response regions overlap. In the bottom panel, we present the response patterns elicited by the first and third stimuli. In this case, peaks are aligned with peaks and valleys with valleys, resulting in small response differences within the overlapping portion of both non-zero response regions. Thus, the magnitude of the response differences within the overlapping portion is much bigger in the case of the first and second stimuli than in the case of the first and third stimuli.
On the neural correlates of psychological spaces

Figure 5.10: Spaces derived through non-metric MDS from the dissimilarities measured at each stage of the Standard Model for four different stimulus sets. The spaces are arranged by model stage across rows and by stimulus set across columns. The stimulus sets are, by order of appearance, line in rectangle stimuli, E’s rotated counterclockwise, L-shapes with middle line varying in angle with respect to the horizontal line segment, and again the line in rectangle stimulus set but with stimuli assuming random positions in the image.

On the other hand, there are fewer units in the non-overlapping portion of the non-zero response regions for the first and second stimuli than for the first and third stimuli. Thus, the magnitude of the response differences within the non-overlapping portion is smaller for the first and second stimuli than for the first and third. Nonetheless, this is not sufficient to compensate for the differences within the overlapping portion. In conclusion, the response differences are more pronounced for the first and second stimuli than for the first and third, i.e. the representations of the first two stimuli are more dissimilar than the representations of
the first and third stimuli. This is in clear contradiction with the relation between the stimuli in the distal space, where the first two stimuli are less distant than the first and third. Whereas we focused on only those units spread along a row of the retinotopic map, the same conclusion is extendible to the remaining units in the map.

*Figure 5.11*: Responses of units at some of the retinotopic maps at stage S1 of the Standard Model to the presentation of a vertical line segment. The retinotopic maps share the same preferred orientation (90°) and correspond to receptive field sizes 7, 17, 27, and 37. The retinotopic maps are presented from left to right by increasing receptive field size.

*Figure 5.12*: Responses of units disposed along the most central row of the retinotopic map corresponding to receptive field size 37 and orientation 90° in S1 stage of the Standard Model to stimuli in the single vertical line segment stimulus set. a) responses to the first and second stimuli; b) responses to the first and third stimuli. In both panels, the response difference produced by each unit is highlighted by a vertical dashed line connecting both responses.
Notice that we confined our analysis to the case of a single retinotopic map. The model, however, contains many other retinotopic maps corresponding to units with different receptive field sizes and preferred orientations. In particular, different receptive field sizes induce different spacings between the peak responses in the non-zero response region, a consequence of the decrease of spatial frequency of the Gabor filters with receptive field size. As a result, violations to monotonicity induced by one retinotopic map are likely to be compensated by the response differences obtained within other retinotopic maps sharing the same preferred orientation. This is the case in the present example. Furthermore, the retinotopic maps corresponding to other preferred orientations produce smaller response differences because of the larger mismatch between their preferred feature and the contents of the stimuli. Consequently, their contribution to the overall response differences is considerably smaller. Nevertheless, they also produce increasing dissimilarity for stimuli at increasing distal distance because of the increase of the portion of non-overlap between the non-zero response regions. In sum, although the Standard Model represents the stimuli at the entry stage in a different manner from that of the Chorus of Prototypes, it is also able to obtain monotonicity between the distal distances and the representation dissimilarities at S1 for this particular stimulus set.

In the case of the subset of the line in rectangle stimuli, the model is no longer able to show monotonicity between representation dissimilarities in S1 and distal distances. In Figure 5.13 we present the representations elicited by the first, middle and last stimuli in the retinotopic maps of units with receptive field size 17. The representation of the inner line segment in the first and last stimuli is partially subsumed by the representations of the leftmost and rightmost borders of the rectangle, respectively, across retinotopic maps. The same does not hold for the inner line segment in the middle stimulus. There are thus more units responding to the line segment in the middle stimulus than units responding to the line segment in the first and last stimulus. At the same time, the units responsive to the line segments in either the first or last stimulus do not respond substantially stronger than those responsive to the line segment in the middle stimulus. This means that when considering the responses of all units in any of the retinotopic maps, the differences between representations is larger for the first and middle stimuli than for the first and last stimuli. This holds for all retinotopic maps with receptive field sizes between 7 and 21 pixels. Since units having the smallest receptive field sizes produce the largest response differences, they contribute the most to the response differences within S1. The result is a clear violation of monotonicity between the dissimilarities at S1 and the distal distances. The space derived through non-metric MDS from the representation dissimilarities at S1 reflects these violations. It is two-dimensional and the stimuli form a C-like configuration with the first and last stimuli at the extreme positions and the remaining ones disposed in between according to distal order. As in the case of the Chorus of Prototypes, the inability to recover a good replica of the distal space at the first stage of the model is mainly due to the normalization of incoming activity at the single unit level. We recalculated the responses of all S1 units under no normalization. The one-dimensional space derived through non-metric MDS from the representation dissimilarities presents a fair stress value (.091) and is, therefore, a good approximation of the stimulus representation in S1. The stimuli are arranged in the space derived in terms of the distal order, although the spacing between stimuli differs between pairs, being considerably smaller between the three most central stimuli than between the remaining pairs.

C1 units respond with the maximum of the responses of all S1 units falling within a contiguous portion of a retinotopic map, which is smaller than the S1 units’ receptive field size and larger than half of it. The nature of the response and the number of S1 units
encompassed by single C1 units induces changes in the manner in which lines in the input image are represented in C1 with respect to S1. The “wavy” pattern obtained in S1 is substituted by a non-zero response region with a single flattened peak at the center and step-wise decaying values towards the sides. The non-zero response region is also wider than the corresponding one in S1 due to the larger receptive field sizes of C1 units. Notwithstanding these changes, the relation between dissimilarity and distal distances for the subset of the line in rectangle under consideration is identical to the S1 counterpart. In Figure 5.14, we present the representations elicited by the first, middle and last stimuli on all the retinotopic maps in band 1.

Figure 5.13: Responses of units in some of the retinotopic maps at the S1 stage of the Standard Model to the presentation of stimuli in the line in rectangle stimulus set. The retinotopic maps share the same receptive field size (17) and differ in preferred orientation (0º, 45º, 90º, and 135º) from top to bottom. The responses to the first, 4th, and 7th stimuli are presented from left to right. The whiter the color, the higher the response of the respective unit.

The same observations made in the case of the S1 representations apply in here. Most importantly, the representation of the inner line in the first and last stimuli is partially subsumed by the representation of the borders of the rectangle while the representation of the inner line in the middle stimulus is not. This is more pronounced in the retinotopic map
whose units are selective to the vertical orientation. Thus, the representations of the first and last stimuli are more similar to each other than the representations of the first and middle stimuli within band 1. This is also the case in bands 2 and 3 with the exception of those retinotopic maps selective to the horizontal orientation. In higher bands, the first and last stimuli are sometimes more dissimilar than the first and middle stimuli. However, the response differences in the lower bands are more pronounced and, therefore, the dissimilarities calculated over all units in C1 present the same properties as those enounced.

![Figure 5.14](image)

*Figure 5.14: Responses of units in the retinotopic maps in band 1 of the C1 stage of the Standard Model to the presentation of stimuli in the line in rectangle stimulus set. The retinotopic maps are presented from top to bottom by increasing angle of preferred orientation (0°, 45°, 90°, and 135°). The responses to the first, 4th, and 7th stimuli are presented from left to right. The whiter the color, the higher the response of the respective unit.*

In the space derived through non-metric MDS from these dissimilarities, the stimuli assume the same configuration obtained from the dissimilarities in S1 (see Figure 5.10). In other words, the representation in C1, although different, preserves the relations between the stimulus representations in S1.

In the next stage of the model, S2b, the stimulus is represented by a multitude of retinotopic maps within every band, each associated with a different (complex) feature. Units at the same position across S2b retinotopic maps are connected to the same set of C1 units and share, therefore, the same receptive field. Their responses vary inversely proportionally to the distance between the respective preferred feature and the response pattern elicited on the C1 units by the stimulus part falling within the receptive field. In a sense, they perform a triangulation of the C1 response pattern elicited by taking the features associated with the S2b units as reference points. Provided the features are sufficiently widespread through the space
of all possible C₁ response patterns, the differences in responses of these S₂b units vary proportionally to the differences between the incoming C₁ response patterns. In short, the relation between C₁ responses tends to be preserved locally in S₂b. Because the dissimilarity between stimulus representations varies proportionally to the accumulated differences measured locally, the local preservation just mentioned guarantees the preservation of stimulus dissimilarity from C₁ to S₂b. We tested this for the current stimuli by deriving a space through non-metric MDS from the stimulus dissimilarities at S₂b. The space is two-dimensional and the stimuli form a C-like configuration just as in the case of C₁. Again, the first and last stimuli appear at the extremes of the configuration with the remaining stimuli appearing in between according to distal order. The inflection at the extremes is, however, less pronounced than the one found for C₁, probably a consequence of the relatively small number of features considered (see Figure 5.10).

In C₂b, the subsequent stage in the model, there are as many units as the number of unit types in S₂b. Each unit responds with the maximum response among all S₂b units sharing the same feature. These have receptive fields centered at different positions in the input image and are selective to the same feature at different scales (bands). Consequently, the representations in C₂b are completely position invariant and moderately scale invariant. This comes along with a potential change to the relations between stimulus dissimilarities with respect to those in S₂b. In S₂b, the stimuli are compared with each other in terms of their contents at each retinotopic position. In the case of the present stimulus set, and because we assumed the stimuli to be aligned by the lower-left vertex of the rectangle, the left border in one stimulus is compared with the left border in any other stimulus, the right border with the right border, etc. In C₂b this is not necessarily the case. Assume a unit type whose feature matches the portion of the first stimulus containing the left border and the inner segment. Let us consider the responses of the S₂b units in a retinotopic map associated with this unit type. In the case of the first stimulus, the maximal response is given by the unit whose receptive field covers the left border and the inner segment. In subsequent stimuli, the inner segment is further to the right and, therefore, there is no unit responding as high as for the first stimulus. The highest responding units have receptive fields covering the left or right border of the rectangle. In the case of the last stimulus, and because the inner segment is at the same distance from the right border as it is from the left border in the first stimulus, the maximal response is given by the unit whose receptive field covers the segment and the right border. The corresponding C₂b unit responds maximally for the first and last stimuli and slightly less for all the remaining stimuli. More importantly, its responses result from matching the same feature with parts of a stimulus at different locations with respect to the overall stimulus.

Different unit types have different feature sizes that can match parts of the stimuli at different scales. This means that the receptive fields of the best matching S₂b units have different sizes, some encompassing only part of a stimulus, some being larger than a stimulus. The representations in C₂b are, thus, a function of the similarity between stimulus subparts of different sizes, some eventually encompassing the full stimulus. The more identical the subparts of two stimuli, the more similar their C₂b representations become. In the case of the present stimulus sets, this does not result in substantial changes to the relation between dissimilarities. The space derived through non-metric MDS from the dissimilarities in C₂b is again two-dimensional. The configuration formed by the stimuli still resembles the C-shape found in S₂b, with the first and last stimuli at the extremes and closer to each other than to the middle stimulus (see Figure 5.10). Eventually, a perfect C-shape would be recovered if more than the 240 features considered were used.
The resemblance between the spaces derived from S2b and C2b dissimilarities might seem surprising at first, taking into account the different ways in which the stimuli are compared with each other in S2b and C2b. However, the example presented above already suggests that the first and last stimulus, for example, have more small subparts in common than the first and middle stimuli. Thus, the dissimilarities resulting from considering smaller subparts present characteristics identical to those identified in the case of the full stimuli, which are already present in S2b. The end product is a set of dissimilarities almost equivalent to that in S2b.

The last stage in the Standard Model coincides with the last stage in the Chorus of Prototypes. There are as many units as stimuli in the stimulus set, each selective to a different stimulus and connected to all units in the previous stage. As in the Chorus of Prototypes, this tends to preserve the relations between stimulus dissimilarities present in the previous stage. The configuration of the stimuli in the space derived through non-metric MDS from the dissimilarities in S4 is identical to the one derived from the dissimilarities in C2b (see Figure 5.10).

In conclusion, the inability of the Standard Model to recover a good replica of the psychological space is mainly a product of unit level normalization taking place at the first stage and whose effects are replicated by the stimulus part similarity matching realized by units in C2b.

We also simulated the Standard Model with the two other stimulus sets having a moving stimulus part with respect to the overall stimulus. The analysis above suggests that the stimulus configuration derived from the dissimilarities at each stage of the model should assume a C-like shape. In Figure 5.10, we present the configurations obtained for both stimulus sets across model stages. In every case, the stimulus configuration does resemble the C-like shape expected with the first and last stimuli at the ends and the middle stimulus the furthest away from these two. Importantly, the stimuli reside in a two-dimensional space whereas the distal space is one-dimensional.

In all simulations reported in this section, we have considered the stimuli to be aligned by their outermost parts at the input image. This allowed us to explain the way in which the representations of the stimuli become related to each other at the latest stages of the model by considering the hierarchical fashion into which these are built up. In principle, the Standard Model should produce exactly the same representations irrespectively of stimulus position. In this particular case, although the first stages are still responsible for providing the similarity measure used in C2b to match stimulus parts, the differences between the full representations of the stimuli at first stages should reflect some effects induced by stimulus transposition. We tested this by assigning the line in rectangle stimuli to random positions in the input image. For reasons that we explore below, we only allowed the vertical and horizontal position to vary in steps of 9 pixels. We used the same model considered in the original condition. The spaces derived from the dissimilarities at each stage are presented in Figure 5.10. In the first three spaces, corresponding to stages S1, C1, and S2b, the stimuli assume a configuration close to the one obtained by assigning each stimulus to the coordinates of its lower left-vertex in the input image. In other words, stimulus dissimilarity at these stages is mainly dependent on the relative position of the stimuli in the input image. In the remaining spaces, corresponding to stages C2b and S4, this is no longer the case. The configuration tends to reflect the major characteristic identified so far for this stimulus set: the first stimuli are at progressively larger distances from the first one and those from the middle stimulus onwards are at progressively smaller distance of the first one. This is only violated for the last stimulus that is at approximately the same distance to the first as the 6th. Contrary to what was
expected, the configuration in the last stage is not identical to the one obtained when the stimuli were presented at the same position. The model seems to fall short in guaranteeing position invariance. The underlying cause is found at the C1 stage. For efficiency reasons, the retinotopic maps in each band are downsampled by considering only the responses of those units at some regular interval from each other. In the case of band 1, the interval is 9 positions wide meaning that a transposed stimulus in the input image only assumes an identical transposed representation if the horizontal and vertical transpositions are multiples of 9 pixels. This justifies the 9 pixels step we have considered. The interval is larger for subsequent bands, increasing in steps of 2. As a result, an identical transposed representation is only obtained in C1 if the stimulus is transposed by the least common multiple of all interval sizes, a value way beyond the size of the input image considered. In conclusion, transposing a stimulus changes the nature of its representation in the model, affecting the dissimilarity between the representations even at the latest stages, a problem that might be overcome by filtering the representation with an anti-aliasing filter before downsampling takes place.

Discussion

Edelman and colleagues (Cutzu & Edelman, 1998; Edelman 1995, 1998; Sugihara et al., 1998) have argued that a view-based theory of object representation is capable of explaining the emergence of psychological spaces at the level of IT. Although this is true for a variety of stimulus sets, we have shown that this does not always hold by addressing the case of two-dimensional stimuli. A mismatch is obtained whenever variation along a psychological dimension is accompanied by substantial changes in the relative position of stimulus parts. The cause is traceable to the first stages of processing where nearby stimulus parts induce non-linear changes in representation, which are preserved throughout subsequent stages. This is not specific to a particular instantiation of the view-based theory as both models considered, the Chorus of Prototypes and the Standard Model, suffer from this.

We have presented three different stimulus sets for which we claim the models are unable to recover the psychological space. In the case of the line in rectangle stimuli, we do have an empirical psychological space to justify our claim. This is no longer the case for the other two stimulus sets. Nevertheless, we do expect the respective psychological spaces to be identical to this one as the variation in the stimuli presents the same characteristics as those found in the line in rectangle stimuli. This awaits confirmation from a behavioral study carried out with human subjects.

In both models considered, visual information flows in a feedforward fashion from the first to the last stage of processing. Brain areas in the ventral visual stream are known to be connected both in a feedforward and feedback manner. Feedback processing is thought to be relevant for “view with scrutiny” (Desimone & Duncan, 1995; Duncan, 1998). Within each processing stage, cells compete for activation. In the absence of external or internal biases, the full contents of the image are represented in a “blurred” fashion across areas. After the first feedforward sweep, some cells in IT will have a small advantage over the others. These will bias the competition at V4, enhancing those cells that contribute the most to their activation. The effects propagate all the way back to V1 and forth. After a while, some items will be represented in a detailed manner whereas the remaining ones are represented with less detail. Such a process could eventually allow the representations in IT, corresponding to the latest stage of the models considered, to match the psychological spaces. In the case of the
line in rectangle stimuli, it would suffice to have the inner line segment enhanced and the leftmost and rightmost borders of the rectangle suppressed. This would reduce or even eliminate the interference between the representation of the inner line segment and the representation of any of the vertical borders, the major obstacle in the recovery of the psychological space. This could be tested empirically through a neurophysiological study. The activity of a set of visually responsive IT cells is collected while a monkey observes passively each stimulus. The average firing rate of a cell is calculated separately for the early and late components of the stimulus induced activity. A space is derived through non-metric MDS from the Euclidean distances between the early average firing rates across the population of cells. Another space is derived in the same way from the late average firing rates. If the hypothesis holds, the first space matches the one derived from the representation of the stimuli in the latest stage of the models while the second matches the psychological space. If these results could be reproduced by an extension of the model by integrating feedback connectivity, then we would conclude that the mismatch we identified is not a limitation of a view based theory but simply a consequence of the non-inclusion of feedback processing.

The considerations above raise a possibility that has been overlooked in the psychological literature. The psychological space might differ depending on whether the data from which it is derived are collected under fast-paced or self-paced conditions. This is not expected for every stimulus set but might be the case for the stimulus sets identified. This hypothesis is not contradicted by the evidence so far, as the psychological space for the line in rectangle stimuli has been derived under self-paced conditions (Erickson & Kruschke, 1998; Kruschke, 1993). A fast-paced condition could be realized by presenting the stimuli for very short intervals (in the order of the tens of milliseconds) and having subjects respond as fast and as accurate as possible. In order to minimize the effects of feedback processes, a backward-masking protocol should be used where a mask is presented shortly after stimulus offset (Serre, Oliva, & Poggio, 2007).

Possibly, the psychological spaces derived under fast- and self-paced conditions are identical. This would suggest that the effects of feedforward processing alone and feedforward and feedback processing together are equivalent in the sense that they induce the same between-stimulus relations. The results presented in here show that this is not the case if the psychological spaces are to be equated with patterns of activity in IT. However, they do not discard the possibility of this taking place at higher brain areas such as the prefrontal cortex (PFC). The PFC is directly interconnected with IT (Ungerleider, Gaffan, & Pelak, 1989; Webster, Bachevalier, & Ungerleider, 1994). Cells in the PFC emphasize behaviorally relevant information such as stimulus category or match/non-match status of the categories of two distinct stimuli (Freedman et al., 2003). In addition, they can be highly selective to small differences between visually and non-visually presented stimuli, especially if these are relevant for the task at hand (Romo et al., 1999; Rainer & Miller, 2000). In the case of the line in rectangle stimuli, assume a set of PFC cells that would emphasize the leftwards and rightwards position of the inner line segment with respect to the center of the rectangle. These cells would produce larger activity differences for those stimuli whose inner line segment is closer to any of the vertical borders of the rectangle than to those having the inner line segment close to the middle. These cells along with a set of cells “reproducing” the responses of cells in IT produce activity patterns that can be shown to be in close correspondence with the psychological space. A neurophysiological study could assess the viability of such an hypothesis.
In conclusion, the viability of a view-based theory in explaining the emergence of psychological spaces is still an open question. In here, we have presented some evidence that suggests the need for further empirical studies to clarify the eventual limits of such a conception.