Learning posture invariant spatial representations through temporal correlations

M. W. Spratling
Division of Engineering, King's College London, London, UK.

Abstract

A hierarchical neural network model is used to learn, without supervision, sensory-sensory coordinate transformations like those believed to be encoded in the dorsal pathway of the cerebral cortex. The resulting representations of visual space are invariant to eye orientation, neck orientation, or posture in general. These posture invariant spatial representations are learned using the same mechanisms that have previously been proposed to operate in the cortical ventral pathway to learn object representation that are invariant to translation, scale, orientation, or viewpoint in general. This model thus suggests that the same mechanisms of learning and development operate across multiple cortical hierarchies.

Keywords: coordinate transformations; invariant representations; dorsal pathway; cerebral cortex; neural networks; epigenetic robotics; sensory-motor mapping

1 Introduction

Visually guided behavior is most naturally defined relative to a number of distinct reference frames or spatial coordinate systems. For example, a saccade takes place within a retina-centered reference frame, reaching for an object takes place in a body-centered frame, while manipulating one object relative to a reference object is performed in a coordinate system centered on the reference object. To support such visually-guided behavior, spatial information is represented in multiple coordinate systems along the dorsal pathway of the cortical visual system (Battaglia-Mayer et al., 2003): neural representations are arranged in a retinotopic map in primary visual cortex (V1), while regions of the parietal cortex contain representations of space in head-centered (Andersen et al., 1985; Duhamel et al., 1997), body-centered (Brotchie et al., 1995), object-centered (Chafee et al., 2007) and world-centered (Snyder et al., 1998) coordinates. A head-centered reference frame could be considered as a representation of retinal position that is *invariant* to eye movements. Similarly, a body-centered representation of visual space is one that is invariant to both eye and neck movements, and an object or world centered reference frame is invariant to eye, neck and body movements. This insight suggests a mechanism by which a hierarchy of coordinate systems (see Fig. 1) could be learned in the dorsal pathway in a manner analogous to that believed to underlie the learning of object representations invariant to viewpoint in the ventral pathway.

Along the cortical ventral pathway, neurons learn to respond to recurring or behaviorally relevant patterns of pre-synaptic activity generated by the neurons in more peripheral cortical regions from which they receive their inputs (Barlow, 1995; Karni, 1996; Karni et al., 1995; Logothetis, 1998; Petersen et al., 1998; Sigala and Logothetis, 2001; Sigman and Gilbert, 2000; Wallis and Bülthoff, 1999; Walsh et al., 1998). Higher-level perceptual representations are thus learned from lower-level ones, and this process can be repeated hierarchically, such that at each stage neurons learn increasing *specialization* together with increasing *invariance*. The viewpoint invariance of these representations results from learning associations across time (Cox et al., 2005; Sakai and Miyashita, 1991; Sinha and Poggio, 1996; Stone, 1998; Stryker, 1991; Wallis, 1998, 2002; Wallis and Bülthoff, 2001). By learning to associate images whose appearance is closely temporally correlated, cortical neurons exploit the fact that objects are generally observed for periods of time, during which they may undergo a number of transformations or be observed from a number of viewpoints. By learning to associate images of an object, seen from different viewpoints, an invariant representation can be formed. This mechanism has formed the basis for a large number of algorithms that learn viewpoint invariance from sequences of images (Bartlett and Sejnowski, 1998; Becker, 1999; Einhäuser et al., 2005; Földiák, 1991; Körding and König, 2001; O'Reilly and McClelland, 1992; Rolls and Milward, 2000; Spratling, 2005; Stone and Bray, 1995; Stringer and Rolls, 2000; Templeman and Loew, 1989; Wallis, 1996; Wallis and Rolls, 1997; Wiskott and Sejnowski, 2002). Hence, a standard method for learning perceptual representations that are invariant to viewpoint is to form temporal associations across image sequences showing object transformations.

The same mechanism of learning temporal correlations between sequences of images can also be used to learn spatial representations invariant to posture. For example, retinocentric visual information can be combined with information about eye orientation to generate a representation of visual space in head-centric coordinates. Such a recoding can be learned since as the eye moves, stationary objects will move across the retina. By recording

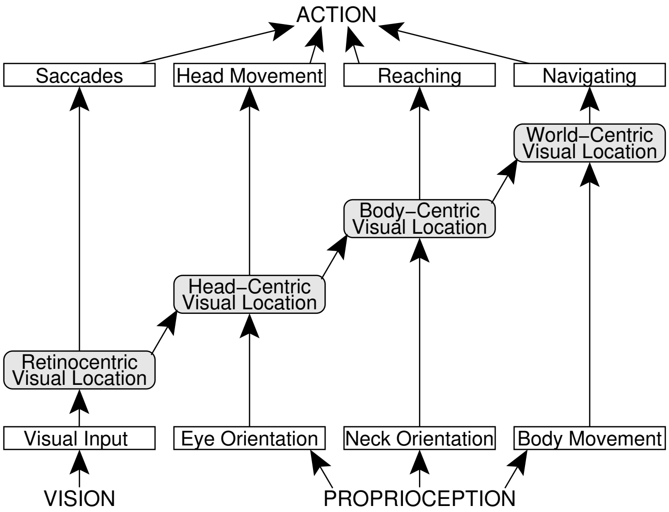


Figure 1: A hierarchy of coordinate systems. It is proposed that such a hierarchy is learned by the cortical dorsal pathway and that each level in the hierarchy is learned by increasing the invariance to posture, *e.g.*, the head-centric reference frame is learned by forming spatial representations invariant to eye orientation, and the body-centric reference frame is learned by forming spatial representations invariant to neck orientation. Each coordinate system is most appropriate for controlling different types of action, however, this article only considers the learning of sensory-sensory coordinate transformations, and *not* how to map those sensory representations to actions.

temporal correlations across sequences of retinal images, it is possible to learn all the combinations of eye orientation and retinal location that correspond to the same point in space (*i.e.*, to learn a spatial representation that is *invariant* to eye orientation). Using the same process during head movements to learn associations between corresponding head-centered visual locations and neck orientations, results in a representation of visual space in body-centric coordinates.

Rather than learning such sensory-sensory coordinate transformations they could be hard-coded using standard mathematical equations for describing kinematics or using neural networks with predefined weight values (see Discussion). The motivation for learning sensory-sensory transformations is two-fold. Firstly, to contribute to research in developmental robotics which seeks mechanisms for creating more adaptive and more autonomous robots as well as mechanisms for creating complex robot control systems that are too complex to design by hand (Spratling, 1999a; Weng et al., 2001). Secondly, to provide a model that may be consistent with the developmental process that occurs during infancy, and hence, which could potentially offer insights into biological development.

2 Methods

Different mathematical processes are required to learn more specialized representations and to learn more invariant representations (Riesenhuber and Poggio, 1999; Spratling, 2005). A more specialized representation results from a node responding to a combination of co-active lower-level features. A node must thus learn to represent a *conjunction* of pre-synaptic inputs. To respond to a conjunction of inputs, a standard weighted sum of pre-synaptic activation values can be used. Such a function will cause the output of the node to be a maximum when all the lower-level features to which it responds are simultaneously active. In contrast, a more invariant representation results from a node responding to multiple, non-coactive, lower-level features. A node must thus learn to represent a *disjunction* of pre-synaptic inputs. To respond to a disjunction of inputs, the maximum of the weighted pre-synaptic activation values can be used. Such a function enables a node to respond invariantly across a number of inputs while maintaining the feature specificity of its response. Hence, several existing architectures for invariant object recognition (Fukushima, 1980, 1988; LeCun et al., 1998, 2004; Riesenhuber and Poggio, 1999; Serre et al., 2002, 2007; Spratling, 2005) consist of alternating layers of neurons that perform these two operations in order to form more specialized representations in one layer, and more invariant representations in the next layer. It has been proposed (Fukushima, 1988; Riesenhuber and Poggio, 1999; Serre et al., 2007) that these two forms of processing correspond to the functionality of simple and complex cells observed in the primary visual cortex (Hubel and Wiesel, 1962).

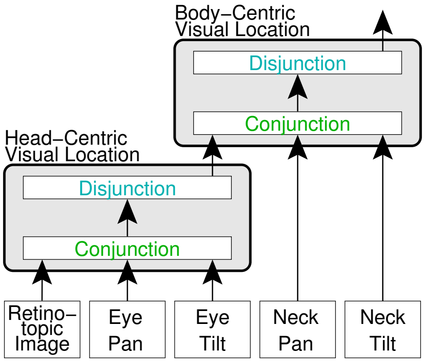


Figure 2: The neural network architecture used to learn a hierarchy of coordinate systems. Two layers of neurons, one learning conjunctions and the other learning disjunctions, are used to perform each sensory-sensory coordinate transform.

The architecture proposed here for learning a hierarchy of spatial coordinate systems also consists of alternating layers of neurons that learn conjunctions and disjunctions (see Fig. 2). One pair of conjunctive and disjunctive layers learns to transform retinocentric coordinates into head-centered coordinates. A second pair of conjunctive and disjunctive layers then learns to transform these head-centered coordinates into body-centered coordinates. The inputs to the hierarchy are presumed to come from more peripheral thalamic regions that are not explicitly modeled. Details of the algorithms employed in the conjunctive and disjunctive layers are provided in the following sub-sections. However, in brief these learning methods are as follows. Each conjunctive layer employs a form of competitive learning that causes a distinct node to learn to represent each distinct input pattern (*e.g.*, for the first conjunctive layer, each distinct combination of eye pan value, eye tilt value and the location of an active pixel in the retinotopic input). Each disjunctive layer employs a form of temporal associative learning that causes a node to learn to represent sets of nodes in the conjunctive layer that are frequently active in sequence (but not simultaneously). If we consider a world containing a single, stationary, object then as the eyes move distinct combinations of eye pan/tilt and retinal input will be generated activating different conjunctive nodes. This sequence of activity in the first conjunctive layer will be learned by a single node in the first disjunctive layer resulting in a single disjunctive node representing all the conjunctive nodes that represent the location occupied by the object. The performance of this algorithm has been tested, as detailed in the Results section, on a simple task in which objects are represented as single pixels in the retinal input. Future work aims at determining if this entirely unsupervised learning method will scale-up to learning sensory-sensory coordinate transformations in more realistic tasks.

2.1 Conjunctive Learning

Each conjunctive layer employs the algorithm proposed in (Spratling et al., 2009). This is an unsupervised, competitive, learning algorithm in which nodes compete to represent unique combinations of inputs. This learning algorithm has been shown to reliably and accurately learn distinct input patterns or image components (Spratling et al., 2009). In this algorithm nodes compete to respond to the current pattern of input activity in a manner that is closely related to a number of other algorithms. This mechanism of competition can be interpreted as: a sequential version of non-negative matrix factorization (Hoyer, 2004; Lee and Seung, 1999), as a divisive form of lateral inhibition that targets the inputs to a population of competing nodes (Harpur, 1997; Harpur and Prager, 1994; Spratling, 1999b; Spratling and Johnson, 2001, 2002), as a neural implementation of Bayes' theorem (Spratling et al., 2009), as a form of divisive normalization like that proposed by Heeger (1991, 1992) and a hierarchy of such networks can be interpreted in terms of both biased competition and predictive coding (Spratling, 2008).

For each conjunctive layer, the output ($\mathbf{y}$) generated in response to an input stimulus ($\mathbf{x}$) was calculated by iteratively updating the following equations:

$$\mathbf{e} = \mathbf{x} \oslash \left(\epsilon + \hat{\mathbf{W}}^T \mathbf{y}\right) \tag{1}$$

$$\mathbf{y} \leftarrow (\epsilon + \mathbf{y}) \otimes \mathbf{W}\mathbf{e} \tag{2}$$

Where $\mathbf{y} = [y_1, \dots, y_n]^T$ is an n-element vector of output activations, $\mathbf{x} = [x_1, \dots, x_m]^T$ is an m-element vector of input activations, $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_n]^T$ is an n by m matrix of weight values, each row of which contains the weights received by a single node, $\hat{\mathbf{W}} = [\hat{\mathbf{w}}_1, \dots, \hat{\mathbf{w}}_n]^T$ is a matrix representing the same synaptic weight values as $\mathbf{W}$ but such that the rows of $\hat{\mathbf{W}}$ are normalized to have a maximum value of one, $\mathbf{e}$ is the inhibited value of the input (or, equivalently, the reconstruction error; see below), and $\oslash$ and $\otimes$ indicate element-wise division and multiplication respectively. The parameter ϵ is a small constant (*i.e.*, 1×10^{-9}) that has a negligible effect on the calculation of $\mathbf{e}$ and $\mathbf{y}$ except to prevent division-by-zero errors when the values of $\mathbf{y}$ are zero. The steady-state values of the node outputs were calculated using 100 iterations of the above equations while the input was held constant.

The following learning rule was applied to the steady-state node activations:

$$\mathbf{W} \leftarrow \mathbf{W} \otimes \left\{1 + \beta\mathbf{y}\left(\mathbf{e}^T - 1\right)\right\} \tag{3}$$

Where β is a positive constant which controls the learning rate. A value of $\beta = 0.025$ was used in the experiments described in this article. Following learning, weights were clipped at zero to ensure that they were non-negative. Weights were initialized to random values chosen from a Gaussian distribution with mean 0.5 and standard deviation 0.125.

This learning algorithm for the conjunctive layers operates by minimizing the error between the input stimulus ($\mathbf{x}$) and the input that is reconstructed from the node outputs ($\hat{\mathbf{W}}^T\mathbf{y}$). The values of $\mathbf{e}$ indicate the degree of mismatch between the top-down reconstruction of the input and the actual input. When a value within $\mathbf{e}$ is greater than unity, indicating that a particular element of the input is under-represented in the reconstruction, the responses of all output nodes receiving non-zero weights from this under-represented input are increased (via equation 2) and the values of weights connecting the under-represented input with active output nodes are also increased (via equation 3). Both these changes will lead to an increase in the strength with which that element is represented in the reconstructed input, and hence, reduce the value of that element of $\mathbf{e}$ towards one (via equation 1). Similarly, when a value within $\mathbf{e}$ is less than unity, indicating that a particular element of the input is over-represented in the reconstruction, the responses of all output nodes receiving non-zero weights from this over-represented input are reduced (via equation 2) and the values of weights connecting the over-represented input with active output nodes are also reduced (via equation 3). Both these changes will lead to a decrease in the strength with which that element is represented in the reconstructed input, and hence, increase the value of that element of $\mathbf{e}$ towards one (via equation 1). When the value of $\mathbf{e}$ is equal to unity the reconstruction of that element is perfect and the weights stop changing due to the term $\left(\mathbf{e}^T - 1\right)$ in equation 3. For elements that are not active in the input vector, the corresponding elements of $\mathbf{e}$ will be zero and the corresponding weights (for active nodes) will stop changing once they have reached a value of zero. Hence, a weight stops changing value when the top-down reconstruction is perfect (*i.e.*, when $\hat{\mathbf{W}}^T\mathbf{y} = \mathbf{x}$) or when the weight is zero.

This algorithm thus finds non-negative, elementary, components of the training data, and uses these components to encode each stimulus with the minimal loss of information (*i.e.*, with the minimal reconstruction error). Unlike many other competitive learning algorithms (see Spratling and Johnson, 2002, 2004, and the references therein), the proposed algorithm does not pre-specify the number of nodes that are required to encode each stimulus. Instead, the number of active nodes is determined by the number of elementary components that are required to accurately represent the input. Hence, each conjunctive layer can represent multiple, co-occurring, objects.

2.2 Disjunctive Learning

Each disjunctive layer implements an improved version of the algorithm proposed in (Spratling, 2005). This algorithm is similar to previous methods that exploit temporal correlations to learn invariant representations. However, it differs from these previous methods in (a) learning correlations between consecutive image pairs, rather than across longer sequences of images, (b) biasing learning so that coincident inputs will be represented by distinct nodes. These modifications enable learning to succeed when the training environment contains multiple, co-occurring, stimuli (Spratling, 2005). Multiple stimuli might occur when the visual input contains multiple objects or background clutter.

For each disjunctive layer, an n-element vector of output values ($\mathbf{y}$) generated in response to an m-element vector of inputs ($\mathbf{x}$) was calculated using the following equation:

$$\mathbf{y} = \max\left\{\hat{\mathbf{W}} \otimes \check{\mathbf{W}} \otimes \mathbf{X}\right\} \tag{4}$$

Where $\hat{\mathbf{W}} = [\hat{\mathbf{w}}_1, \ldots, \hat{\mathbf{w}}_n]^T$ is an n by m matrix representing the synaptic weight values such that the rows of $\hat{\mathbf{W}}$ are normalized to have a maximum value of one, $\check{\mathbf{W}} = [\check{\mathbf{w}}_1, \ldots, \check{\mathbf{w}}_n]^T$ is an n by m matrix representing the same synaptic weight values such that the columns of $\check{\mathbf{W}}$ are normalized to have a maximum value of one, $\mathbf{X} = [\mathbf{x}^T, \ldots, \mathbf{x}^T]^T$ is a n by m matrix each row of which is a copy of the input vector $\mathbf{x}$, and $\max$ is a function which returns the maximum value in each row.

The activation function, described above, is identical to that used in (Spratling, 2005). The learning rule, described below, is different from that used previously, although it operates using the same principles. The following learning rule was applied:

$$\mathbf{W} \leftarrow \mathbf{W} \otimes \{1 + \gamma \mathbf{y} (\mathbf{x}^* - \mathbf{x})\} \tag{5}$$

Where $\mathbf{x}^*$ is the vector of input values for the *previous* input pattern, and γ is a positive constant which controls the learning rate. A value of $\gamma = 0.25$ was used in the experiments described in this article. Importantly, for each node the single synapse that produced the maximum value in equation 4 did not have its weights modified. Following learning, synaptic weights were clipped at zero (*i.e.*, negative weight values were made equal to zero) and were normalized such that the sum of the weights in each column of $\mathbf{W}$ was equal to one. Weights were initialized to random values chosen from a Gaussian distribution with mean $\frac{0.5}{n}$ and standard deviation $\frac{0.125}{n}$.

This learning algorithm for the disjunctive layers operates by (a) finding correlations between successive input patterns, and (b) ensuring that coactive inputs (which must be generated by the co-activation of distinct retinal locations) are not represented by the same node. Objective (a) is achieved by nodes which are active in response to the current input pattern increasing their synaptic weights to inputs which were active in the previous stimulus (in proportion to $\mathbf{y}\mathbf{x}^*$). A strongly active node has a strong weight to an input that is active in the current stimulus. Hence, by increasing this node's synaptic weights to inputs that were active in the previous stimulus the algorithm is learning temporal correlations between successive stimuli, in a manner consistent with Hebbian learning. Objective (b) is achieved by nodes which are active in response to the current input pattern decreasing their synaptic weights to inputs which are active in the current stimulus in an anti-Hebbian manner (*i.e.*, in proportion to $-\mathbf{y}\mathbf{x}$). Note, only those synapses that did not generate the maximum input to the node are modified, so a single active input remains unchanged but weights to all other active inputs are deceased. Furthermore, the weights are normalized so that the sum of the weights emanating from a single input are equal to one. This normalization process provides an implicit form of competition between nodes, since if one node strengthens its connection to a particular input, then connections from that input to all other nodes are weakened.

3 Results

A two stage hierarchy (as illustrated in Fig. 2) was used to learn to encode sensory input in both head-centric and body-centric coordinates. The inputs to this hierarchy were a two dimensional image, and four one-dimensional arrays encoding the four motor degrees of freedom: eye pan/tilt and neck pan/tilt. The axes of the motor actions were coincident with the x- and y-axes of the retinal array, hence, both the eye and neck pan/tilt values served to slide the retina around within a larger world coordinate system, as illustrated in Fig. 3. This visual world consisted of a grid of pixels large enough to accommodate the retina at the extremes of the pan/tilt ranges. The world was populated by "objects" represented by pixels with contrast values randomly selected within the range 0 to 1. Since real objects have an extent, representing objects as pixels is a limitation of the current algorithm, but one that it shares with other methods for performing sensory-sensory transformation (see Discussion).

Experiments were carried out using a small 3-by-3 pixel retina, and pan and tilt values that could each take one of three values. The world was thus 7-by-7 pixels. For these experiments the first conjunctive layer contained 180 nodes, the first disjunctive layer contained 50 nodes, the second conjunctive layer contained 550 nodes, and the second disjunctive layer contained 100 nodes. Hence, each layer contained approximately twice as many nodes than the minimum number required for the task. While this is a small scale problem, it is still more complex than the task used to test many existing methods (see Discussion) and the network is still quite complex containing 97500 synaptic weights.

The initial world was randomly created with each pixel independently selected to contain an object with probability P_s (P_s sets the sparsity of the visual input). Values of P_s ranging from 0.05 to 0.3 were used in the different experiments reported below. Each selected pixel then had its contrast set to a value chosen uniformly from the range [0,1]. In order to generate a sequence of data in which objects tended to remain stationary for an extended period of time, each subsequent visual world was created by removing, with probability P_{off}, existing objects (by setting the corresponding pixel value to zero). Similarly, new objects could appear at empty pixel locations with probability P_{on}. New objects were assigned a random contrast in the range [0,1]. In order to keep the sparsity of the world constant (at P_s), P_{on} was related to P_{off} by $P_{on} = \frac{P_{off}P_s}{1-P_s}$. Values of P_{off} ranging from 0.05 to 0.3 were used in the different experiments reported below. Worlds that contained no objects were removed from the training sequence.

The retinal input corresponding to each world array was generated by randomly selecting pan and tilt values with the restriction that the highest contrast pixel in the world remained visible on the retina. Each retinal input was then normalized by the total sum of the pixel values in the retina. To generate the data for training the head-centric representation, the neck pan/tilt values were kept constant while the eye pan/tilt values were chosen randomly. To

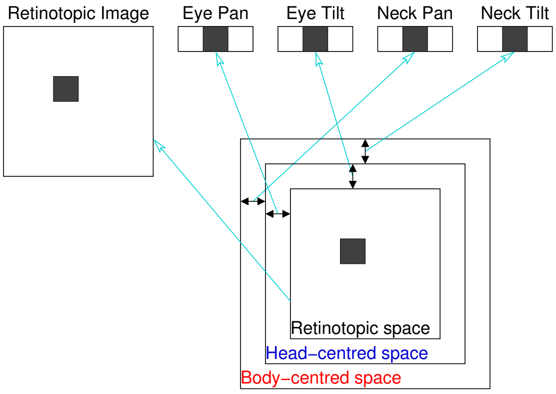
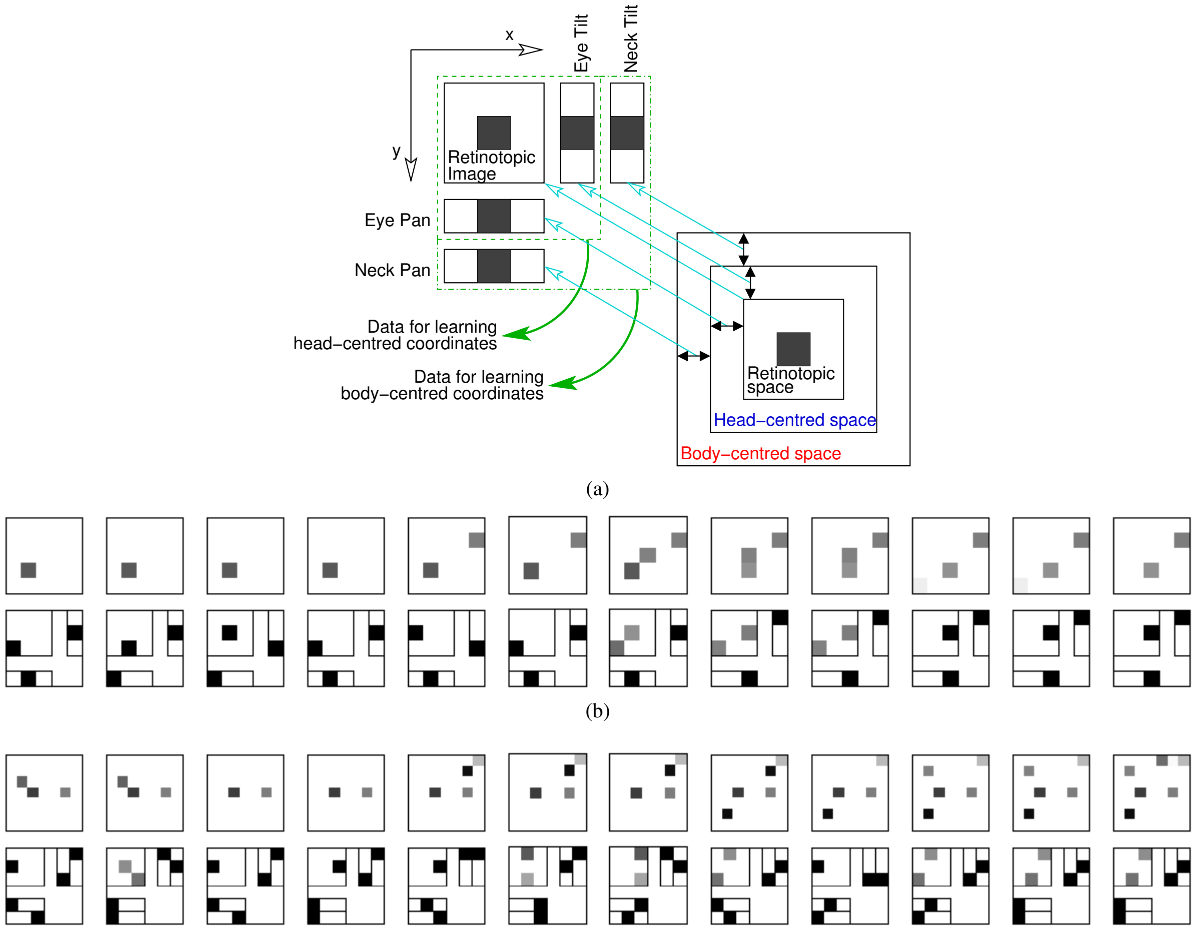


Figure 3: Training data, for presentation to the inputs of the neural network architecture shown in Fig. 2, was generated by populating a body centric world with objects represented by pixels with non-zero values (one object is illustrated near the center of the retinocentric space). The retina was slid around within a world by both eye and neck movements. The portion of the world visible on the retina at the extremes of eye pan and tilt (for a fixed neck pan and tilt) defines a head-centric world. The world visible on the retina at the extremes of both eye and neck pan and tilt defines a body-centric world. Values of pan and tilt were represented to the network as vectors. Each pan/tilt vector had one element equal to a value of one, representing the pan/tilt value, and all other elements were zero. The retinotopic image presented to the network was simply the array of world pixels visible to the retina given the current pan and tilt values. The strength of the pixel contrasts in the retina were normalized so that they summed to one.

generate the data for training the body-centric representation, both the eye pan/tilt and neck pan/tilt values were chosen randomly. Examples of sequences of training data generate by this method are shown in Fig. 4. The four layers of neurons making up the model were trained one after the other. Specifically, a sequence of 100000 input patterns were used to train the first conjunctive layer. These weights were then held constant while another 100000 input patterns were presented and the weights into the first disjunctive layer were learned. This procedure was then repeated to train the second conjunctive and disjunctive layers using 100000 input patterns each.

To test the accuracy of the sensory coordinate mappings learned by the network the following procedure was used. A single object was placed in the visual world at each possible location in turn. For each world location a set of input patterns were created containing every possible combination of the retinal location and the pan/tilt values corresponding to this world location. This set of input patterns corresponding to a single world location constitutes a “disjunctive set” of patterns, and the entire test set consisted of disjunctive sets for every possible world location. For each individual input pattern the node in the disjunctive layer under test that generated the strongest response was identified. This node can be said to represent that input pattern. Ideally, all the patterns in a disjunctive set (*i.e.*, all the combinations of pan/tilt and retinal coordinates corresponding to a single world location) should be represented by a single node, and patterns in different disjunctive sets (*i.e.*, patterns corresponding to different world locations) should be represented by distinct nodes. Hence, a pattern was considered mis-represented if it was not represented by the node representing the majority of patterns in that disjunctive set. Furthermore, if a single node represented patterns corresponding to more than one distinct world location, those patterns not forming part of the largest disjunctive set represented by that node were also considered mis-represented. A percentage error was then calculated as the ratio of mis-represented patterns to all patterns in the test set (*i.e.*, all possible combinations of retinal location and pan and tilt values for all possible world locations). Note that error was *not* calculated in terms of the distance between the actual location of the object and the location represented on the map. Instead, the error values measured the percentage of patterns mis-mapped by the network. In other words, for the percentage of patterns not mis-mapped (100-error) the mapping produced was 100% accurate.

Experiments were carried out to assess the robustness of the learning algorithm to changes in the sparsity of the world (P_s), and the probability that successive images contained the same object (P_{off}). For each condition tested, the neural architecture was trained five times using different randomly generated sequences of input data and different randomly initialized synaptic weights. Fig. 5 shows the mean percentage error (and the maximum and minimum error) over these five trials for each combination of P_s and P_{off} tested. It can be seen from Fig. 5a that performance was best for P_s values between 0.1 and 0.2. As the world becomes sparser than 0.1, the retina will rarely contain multiple objects. Since, the disjunctive learning rule exploits the presence of multiple objects to divide its inputs into separate disjunctive sets, it is not surprising that performance gets worse as P_s becomes very small. As the world becomes denser than 0.2, the retina will typically contain several objects. It is then difficult to

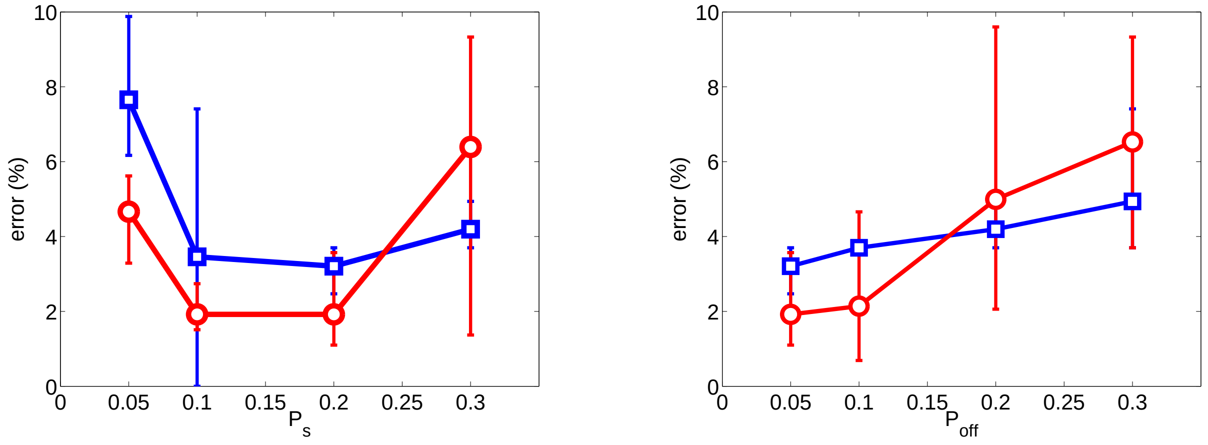


(a)

(b)

(c)

Figure 4: Example training data. (a) shows Fig. 3 rearranged to illustrate the format in which the training data is presented in (b) for learning head-centered coordinates, and (c) for learning body-centered coordinates. In both (b) and (c) the top rows show a sequence of 12 images of a visual world created using $P_s = 0.1$ and $P_{off} = 0.1$. The bottom rows show the corresponding 12 sets of input data presented to the neural network. For learning head-centric coordinates it is assumed that the neck pan and tilt values are constant. Hence, the visual world, shown on the top row of (b), is the head-centered space (fixed relative to the body-centered space) as shown on the right of (a). Furthermore, since the neck pan/tilt values are not used as inputs to the stage of the neural network hierarchy learning the head-centric coordinates (see Fig. 2), these values are ignored. Hence, the sequence of training data shown on the bottom row of (b) consists of the retinotopic image and the eye pan and tilt vectors shown in the configuration illustrated within the dashed box on the top-left of (a). For learning body-centric coordinates the visual world, shown on the top row of (c), is the body-centered space shown on the right of (a). The sequence of training data shown on the bottom row of (c) consists of the retinotopic image and both the eye and neck pan and the eye and neck tilt vectors shown in the configuration illustrated within the dash-dot box on the top-left of (a). In both sets of training data the retina is smaller than the visual world, and hence the retinal inputs (top-left square in the bottom row of each sub-figure in (b) and (c)) contains only a subset of the pixels shown in the corresponding visual world (top row of the same column). The retina is positioned on the visual world at the location determined by the pan and tilt values, and hence, the subset of pixels contained in the retinotopic image is determined by the pan and tilt vectors. For example, for the left most column in (b) both the pan and tilt values are in the middle of their ranges (pan=2, tilt=2) and so the top and left edges of the retina are positioned one pixel from the top and left edges of the world (and the bottom and right edges of the retina are positioned one pixel from the bottom and right edges of the world). Hence, the only object in this world appears at the bottom left of the retina (at coordinates (1,3)). In the next image in the sequence, the eye pan value has moved one position to the left (pan=1, tilt=2) so that the left edge of the retina is now positioned at the left edge of the world and the object now appears in the retina at the bottom center pixel (at coordinates (2,3)). If for brevity we denote each input pattern using a vector of four values (pan,tilt,x,y), then the first five sets of input data for learning head-centered coordinates shown on the bottom row of (b) can be written as (2,2,1,3), (1,2,2,3), (1,3,2,2), (2,2,1,3), (2,3,1,2). Note that the first and fourth patterns are the same, so we have four distinct training patterns. Each of these patterns represents the same location in the visual world (the location of the pixel at coordinates (2,4) in the visual world shown on the top row of (b). These four patterns form a "disjunctive set" representing that location.



(a)

(b)

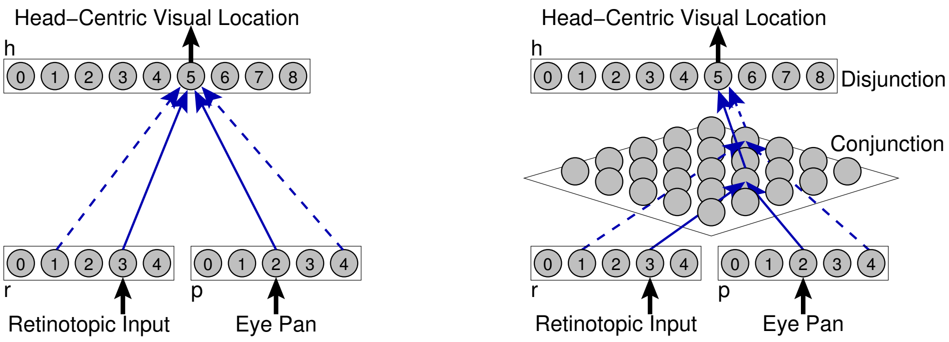
Figure 5: Accuracy of head-centric representation (square markers) and body-centric representation (circular markers), for (a) changing the sparsity of objects in the world (while P_{off} was fixed at 0.05), and (b) changing the average time an object stayed static in the world (while P_s was fixed at 0.2). Each plot shows the mean percentage error with error bars indicating the best and worst performance over five trials using different randomly generated sequences of input data and different randomly initialized synaptic weights.

learn which of several objects in one image corresponds to which of several objects in the next image. Hence, it is to be expected that learning posture invariance from temporal correlations becomes less accurate as the average number of objects in each input pattern increases. However, it is easy to imagine how the early stages of visual processing could control the sparseness of the retinocentric representation of space, to enable successful learning of subsequent sensory-sensory transformations. Fig. 5b shows that the performance of the learning algorithm improves as P_{off} decreases. This is expected since the smaller P_{off} the longer the sequence of training data containing the same object, and hence, the more opportunity the algorithm has to learn all combinations of retinal position and pan/tilt value that correspond to the same spatial location.

It can be seen from both Figs. 5a and 5b that the body-centric representation is often more accurate than the head-centric representation. Since the former is built on the latter, this result seem inconsistent. However, it is due to the rather strict criteria used to quantify the accuracy of the representations learned. The assessment method requires that each disjunctive set is represented by a distinct node, and that a single node produces the strongest response to all members of a disjunctive set. The second criteria of success is responsible for many of the recorded errors, but such "errors" in the head-centric representation do not prevent the body-centric representation being learned successfully. For a small proportion of disjunctive sets two or more nodes in the disjunctive layer learn strong weights to all the members of this disjunctive set. Unless one of these nodes produces the strongest response to all members of the disjunctive set, we classify some of the patterns making up the disjunctive set as being misrepresented. However, if multiple nodes all respond to a disjunctive set, the conjunctive layer in the next stage of the learning hierarchy can learn strong connections to all these nodes and still form an accurate representation of all possible conjunctions at the next level. Hence, some of the errors being counted by the assessment method, when applied to the head-centric representation, do not actually have an effect on subsequent learning of the body-centric representation. Such "errors" seem to be most common at the corners of the space being encoded. Of all the combinations of retinal location and pan/tilt values a smaller proportion are near corners in the body-centric space than in the head-centric space, hence, the accuracy of the body-centric representation is better.

4 Discussion

This article has shown that it is possible to learn sensory-sensory coordinate transformations using a completely unsupervised learning algorithm. Specifically, it has been shown that a sequence of retinal images produced using random eye movements and a slowly changing world can be used to learn a representation of visual space that is invariant to eye movements, *i.e.*, a head-centric reference frame. Similarly, this representation of the visual world combined with random head movements can be used to learn a representation of visual space that is invariant to eye and neck movements, *i.e.*, a body-centric reference frame. Other recent work has shown that similar methods



(a)

(b)

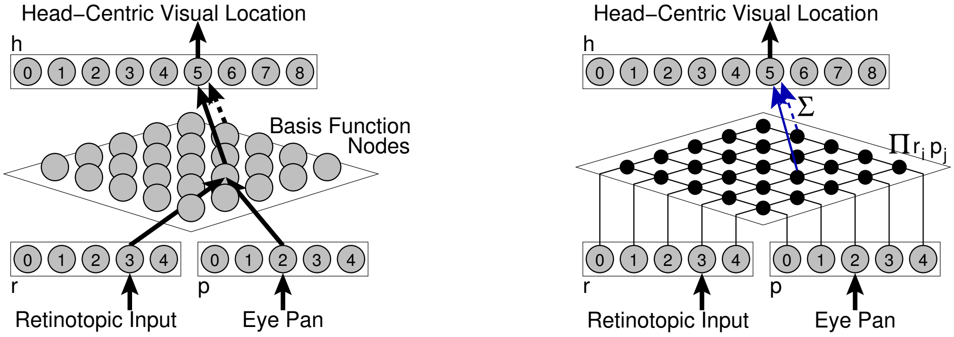
Figure 6: (a) A simple task used to test previous models of sensory coordinate transformations. Two scalar values r_i and p_j, encoded by the activation pattern of two populations of neurons, are combined to calculate a third scalar value h_k, encoded by the firing of a third population of neurons. This general task can be interpreted in terms of simulating a mapping from retinocentric coordinates to head-centered coordinates, if r_i is interpreted as representing the position of a target on a one-dimensional retina, and p_i is interpreted as a representation of eye position. In the simple case where the eye position generates a horizontal shift of the retina along its axis, the correct value of h is simply the sum of the values r and p. Hence, two pairs of retinocentric coordinates and eye position values that should produce the same head-centric output are indicated by solid and dashed lines respectively. (b) The architecture of the proposed model when applied to this simple task. The head-centered representation is generated by a population of disjunctive nodes. These nodes learn strong weights from a set of conjunctive nodes (all representing the same spatial location). The conjunctive nodes learn each possible combination of retinotopic and eye position inputs. The conjunctive layer thus learns a set of basis functions, and the disjunctive layer learns the mapping from these basis functions to a coordinate system in which object position is invariant to posture.

which exploit temporal correlations can be used to learn spatial representations invariant to posture similar to those encoded by "place cells" in the rat hippocampus (Byrne and Becker, 2008; Franzius et al., 2007). The current results complement this work by showing that similar mechanisms can account for learning spatial reference frames in parietal cortex.

The reported results demonstrate that the proposed algorithm works on a simple test problem. Future work will be required to determine if this algorithm can scale-up to more realistic tasks. Existing methods for generating sensory-sensory mappings use hard-coded solutions or supervised learning, and hence, are likely to outperform the method proposed here. The advantages of the current method are not in terms of a quantitative improvement in performance, but rather in demonstrating that a completely unsupervised method can learn sensory-sensory coordinate transformations. Hence, in the following discussion the comparison with other methods is concerned primarily in discussing the shortcomings of these existing methods in terms of their plausibility as models of development in infants and/or as algorithms that can advance work in epigenetic robotics.

The proposed algorithm has been tested with a simple task using a two-dimensional retinal image and four degrees of freedom for motor action: eye pan/tilt and neck pan/tilt. While this task is simple it is more challenging and realistic than the task that has typically been used to assess previous models of cortical spatial transformations. This previous task is illustrated in Fig. 6a. In this task the retinotopic map is one-dimensional and there is one degree of freedom for motor action which causes the retinal input to translate along its length (Pouget and Sejnowski, 1997; Weber and Wermter, 2007). This is equivalent to calculating the sum of two scalar values represented by peaks of activity in two one dimensional input maps (van Rossum and Renart, 2004). For ease of comparison with previous methods, the architecture of the proposed algorithm when applied to this simple one-dimensional problem is shown in Fig. 6b.

Basis function networks (Fig. 7a) are an influential model of cortical spatial transformations (Deneve and Pouget, 2003; Pouget et al., 2002; Pouget and Sejnowski, 1997; Pouget and Snyder, 2000). In such models, a layer of basis function nodes encode every possible combination of sensory input signals. This layer is functionally equivalent to the conjunctive layer in the proposed model (compare Fig. 7a with Fig. 6b). The outputs of the basis functions can be combined, linearly, to produce a spatial map in a new reference frame (which is equivalent to the operation performed by the disjunctive layer in the proposed model). The need to assign one node to represent each combination of sensory signals results in the basis function network size increasing exponentially



(a) (b)

Figure 7: Architectures of alternative methods of performing sensory reference frame transformations for the one-dimensional task described in Fig. 6a. (a) A basis function network as used in (Deneve and Pouget, 2003; Pouget et al., 2002; Pouget and Sejnowski, 1997; Pouget and Snyder, 2000), (b) A sigma-pi network as used in (Weber et al., 2007; Weber and Wermter, 2007).

with problem size (Deneve and Pouget, 2003; Pouget and Sejnowski, 1997). To resolve this issue it is possible to decompose sensory-sensory transformations into several steps. Such a solution results in a model containing a hierarchy of coordinate systems (Pouget et al., 2002) similar to the one proposed in this article. However, basis function networks differ in two key respects from the proposed model. Firstly, because the response of a neuron in a basis function network is inversely proportional to the Euclidean distance between the input and the weights of that node, basis function networks require that all the sensory information originates from a single object (Pouget et al., 2002), and hence, unlike the proposed model cannot represent the spatial locations of multiple objects that are present simultaneously. Secondly, the synaptic weights of each basis function need to be predefined, hence, unlike the proposed model spatial representations are not learned in an unsupervised way, and such a model cannot be used to explore the development of spatial representations in cortex.

Weber et al. (2007); Weber and Wermter (2007) proposed a method for learning sensory reference frame transformations. The nodes in their network were sigma-pi units (Fig. 7b). The response of each node was determined by calculating the products of pairs of signals, and then calculated the sum over these products. The two parts of this activation function correspond to the two layers in the proposed model and a basis function network (compare Fig. 7b with Figs. 6b and 7a). As with the proposed model, the weights connecting products to outputs are learned using temporal correlations between successive input patterns. However, unlike the proposed model, the combinations of inputs that form the products are predefined. This is equivalent to predefining the weights in a basis function network, or predefining the weights in the conjunctive layers of the proposed model. Hence, only part of the mapping is learned in this algorithm, while half the problem is solved by hard-wiring the solution rather than learning it.

Hence, in contrast to the algorithm proposed here (and to similar models of hippocampal spatial representations (Byrne and Becker, 2008; Franzius et al., 2007)), previous models have failed to learn the sensory coordinate transformation but have had the solution hard-wired into the weights of the network to some extent. Furthermore, none have demonstrated that the resulting spatial representation can be used as the input to a subsequent spatial reference frame mapping (*i.e.*, that the model can form part of a larger hierarchy of spatial transformations).

The proposed model, and those discussed above, suggest that when performing a task such a reaching for an object the retinocentric coordinates of the target object are transformed into body-centered coordinates via a series of intermediate reference frames. Hence, sensory information is recoded into a coordinate system which is more likely to be consistent with the coordinate system of the motor effectors. The implicit assumption is that the new coordinate systems is more appropriate for learning and controlling arm movements. However, an alternative strategy, embodied in many existing models, is to directly learn the mapping between sensory inputs and motor outputs without any recoding via intermediate reference frames (*e.g.*, Albus, 1981; Balkenius, 1995; Baraduc et al., 2001; Churchland, 1990, 1986; Cohen et al., 1997; Coiton et al., 1991; Gaudiano and Grossberg, 1991; Mel, 1990; Meng and Lee, 2007; Metta et al., 1999; Ritter et al., 1992; Salinas and Abbott, 1995; Schulten and Zeller, 1996). Unlike many previous algorithms for sensory-sensory mappings, these sensory-motor mappings are usually learned rather than being hard-wired. This is made possible by associating the sensory consequences of random motor movements (motor "babbling") with the outputs that generated those motor movements. In other words, the motor outputs provide a supervisory signal for learning the sensory-motor mapping that is not available for learning a sensory-sensory mapping.

While the ability to exploit the correlation between motor outputs and sensory inputs to learn the sensory-motor mapping gives these methods an advantage over hard-coded methods of learning sensory-sensory mappings, the existing algorithms for learning direct sensory-motor mappings typically suffer from other issues.

- These existing algorithms typically require that the sensory input contain only one item which always corresponds exactly to the position of the end-effector that the algorithm is learning to control, which is a serious limitation.

- During training the output of the motor region is determined by a random number generator, whereas it is controlled by the sensory input once learning is complete. Such a distinction between the training phase and operational phase is biologically implausible as it requires neurons to switch behaviors (Spratling and Hayes, 1998).

- There are a large number of possible combinations of sensory inputs for which a mapping needs to be learned (Tsotsos, 1995), this is equivalent to the problem faced by basis function networks where the size of the network needs to increase exponentially with the size of the task.

- The required motor configuration may depend on non-linear relationships between the different sensory inputs. There may thus be little similarity in the required output for similar input patterns, and hence, little opportunity for generalization.

At least some of these difficulties can be solved by recoding the sensory data into a more abstract representation such as a body-centric reference frame. Recoding via intermediate coordinate systems can overcome the exponential increase in problem size, the more direct correspondence between the recoded sensor and motor coordinate systems provides generalization between similar situations, and the abstract sensory representation can be reused for learning other sensory-motor mappings.

As well as providing a mechanism for learning sensory-sensory coordinate transforms (and subsequently, sensory-motor control) the proposed hierarchical neural network algorithm may have application to modeling infant development or implementing developmental processes in robots. Development is a constructivist process through which progressive improvements in ability are achieved by using simpler skills as a basis for learning more complex ones, and so on hierarchically (Arbib et al., 1987; Elman et al., 1996; Mareschal et al., 2007; Quartz, 1999; Siros et al., 2008). A neural network model of such a process thus requires nodes to be available to exploit the learning that has been achieved by other nodes. One obvious method by which this can be achieved is to allow the outputs of certain nodes to provide (some of) the inputs to other nodes, in a hierarchical arrangement. This second set of nodes is then in a position to make use of, and build upon, the results of learning in first set of nodes. In such a hierarchical neural architecture, more complex representations can be learned in higher-level networks based on simpler representations learned in lower-level networks. In addition, the simpler representations constrain the search space for learning in subsequent layers, and hence, make tractable the task of learning more complex representations (Clark and Thornton, 1997; Elman et al., 1996). The organization of cortical regions into functional hierarchies, such as those found in the dorsal and ventral streams, suggest that hierarchical neural architectures, like that proposed here, can provide the basis for a model of cortical development that is more biologically plausible than many other neural network algorithms, like those reviewed in (Westermann et al., 2006), that have been proposed as models of development. Neural hierarchies capable of learning complex perceptual representations which are appropriate for controlling complex actions are likely to be essential to the developmental process in humans, and algorithms capable of learning a hierarchy of representations are therefore likely to be essential if the field of developmental robotics (Spratling, 1999a; Weng et al., 2001) is to make further progress, which in turn, is imperative for the creation of more intelligent and adaptive machines.

In the proposed model, initial learning of a head-centric representation simplifies the subsequent task of learning a body-centric representation. In turn, the learning of the head-centric spatial representation requires visual inputs generated by eye movements in the absence of neck movements. Hence, the model proposes that the lack of head control in young infants is an advantage for learning spatial representations, in the same way that it has previously been proposed that limitations in motor control or cognitive ability may provide an advantage to learning complex tasks (Araujo and Grupen, 1996; Elman, 1993; Elman et al., 1996; Lungarella and Berthouze, 2002; McClelland and Plaut, 1993). Another prediction of the proposed model is that learning invariant representations will be aided by the presence of multiple stimuli. This is in contrast to other algorithms for learning invariance which require stimuli to be presented in isolation (*e.g.*, Földiák, 1991; Oram and Földiák, 1996; Stringer and Rolls, 2000; Wallis, 1996; Wallis and Rolls, 1997).

5 Conclusions

The structural uniformity of the neocortex (Crick and Asanuma, 1986; Ebdon, 1992; Mountcastle, 1998) has led many theorists to suggest that the cortex is also computationally uniform (Barlow, 1994; Douglas et al., 1989; Ebdon, 1992, 1996; Grossberg, 1999; Marr, 1970; Mumford, 1992, 1994; Phillips and Singer, 1997). This article adds to these arguments by demonstrating that the same mechanisms which are widely believed to underlie the learning of object representations with invariance to viewpoint in the ventral pathway can also give rise to spatial representations invariant to eye and neck movements that are believed to exist in the dorsal pathway. A major advantage of the proposed algorithm over previous models of cortical sensory-sensory transformations is that these coordinate transformations are learned using an entirely unsupervised learning algorithm.

Acknowledgments

This work was funded by EPSRC Research Grant EP/D062225/1.

References

Albus, J. S. (1981). *Brains, Behaviour and Robotics*. BYTE Books.

Andersen, R. A., Essick, G. K., and Siegel, R. M. (1985). Encoding of spatial location by posterior parietal neurons. *Science*, 230(4724):456–8.

Araujo, E. G. and Grupen, R. A. (1996). Learning control composition in a complex environment. In Maes, P., Mataric, M., Meyer, J.-A., Pollack, J., and Wilson, S. W., editors, *From Animals to Animats: Proceedings of the International Conference on Simulation of Adaptive Behaviour*, pages 333–42, Cambridge, MA. MIT Press.

Arbib, M. A., Conklin, E. J., and Hill, J. C. (1987). *From Schema Theory to Language*. Oxford University Press, Oxford, UK.

Balkenius, C. (1995). Multi-modal sensing for robot control. In Niklasson, L. F. and Bodén, M. B., editors, *Current trends in connectionism*, pages 203–16. Lawrence Erlbaum, Hillsdale, NJ.

Baraduc, P., Guigon, E., and Burnod, Y. (2001). Recodning arm position to learn visuomotor transformations. *Cereb. Cortex*, 11:906–17.

Barlow, H. B. (1994). What is the computational goal of the neocortex? In Koch, C. and Davis, J. L., editors, *Large-Scale Neuronal Theories of the Brain*, chapter 1. MIT Press, Cambridge, MA.

Barlow, H. B. (1995). The neuron doctrine in perception. In Gazzaniga, M. S., editor, *The Cognitive Neurosciences*, chapter 26. MIT Press, Cambridge, MA.

Bartlett, M. S. and Sejnowski, T. J. (1998). Learning viewpoint invariant face representations from visual experience in an attractor network. *Network: Computation in Neural Systems*, 9(3):1–19.

Battaglia-Mayer, A., Caminiti, R., Lacquaniti, F., and Zago, M. (2003). Multiple levels of representation of reaching in the parieto-frontal network. *Cereb. Cortex*, 13(10):1009–22.

Becker, S. (1999). Implicit learning in 3D object recognition: the importance of temporal context. *Neural Computation*, 11(2):347–74.

Brotchie, P. R., Andersen, R. A., Snyder, L. H., and Goodman, S. J. (1995). Head position signals used by parietal neurons to encode locations of visual stimuli. *Nature*, 375(6528):232–5.

Byrne, P. and Becker, S. (2008). A principle for learning egocentric-allocentric transformations. *Neural Computation*, 20(3):709–37.

Chafee, M. V., Averbeck, B. B., and Crowe, D. A. (2007). Representing spatial relationships in posterior parietal cortex: Single neurons code object-referenced position. *Cereb. Cortex*, 17(12):2914–32.

Churchland, P. M. (1990). Some reductive strategies in cognitive neurobiology. In Boden, M. A., editor, *The Philosophy of Artificial Intelligence*, chapter 14. Oxford University Press, Oxford, UK.

Churchland, P. S. (1986). *Neurophilosophy*, chapter 10. MIT Press, Cambridge, MA.

Clark, A. and Thornton, C. (1997). Trading spaces: computation, representation and the limits of uninformed learning. *Behavioral and Brain Sciences*, 20(1):57–66.

Cohen, P. R., Atkin, M. S., Oates, T., and Beal, C. R. (1997). Neo: learning conceptual knowledge by sensorimotor interaction with an environment. In *Proceedings of the 1st International Conference on Autonomous Agents*, pages 170–7. ACM Press.

Coiton, Y., Gilhodes, J. C., Velay, J. L., and Roll, J. P. (1991). A neural network model for the intersensory coordination involved in goal-directed movements. *Biological Cybernetics*, 66(2):167–76.

Cox, D. D., Meier, P., Oertelt, N., and DiCarlo, J. J. (2005). 'Breaking' position-invariant object recognition. *Nature Neuroscience*, 8(9):1145–7.

Crick, F. and Asanuma, C. (1986). Certain aspects of the anatomy and physiology of the cerebral cortex. In Rumelhart, D. E., McClelland, J. L., and The PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructures of Cognition. Volume 2: Psychological and Biological Models*, pages 333–71. MIT Press, Cambridge, MA.

Deneve, S. and Pouget, A. (2003). Basis functions for object-centered representations. *Neuron*, 37:347–59.

Douglas, R. J., Martin, K. A. C., and Whitteridge, D. (1989). A canonical microcircuit for neocortex. *Neural Computation*, 1(4):480–8.

Duhamel, J.-R., Bremmer, F., BenHamed, S., and Graf, W. (1997). Spatial invariance of visual receptive fields in parietal cortex neurons. *Nature*, 389.

Ebdon, M. (1992). The uniformity of cerebral neocortex and its implications for cognitive science. Technical Report CSRP-228, School of Cognitive and Computing Sciences, University of Sussex.

Ebdon, M. (1996). *Towards a General Theory of Cerebral Neocortex*. PhD thesis, University of Sussex, UK.

Einhäuser, W., Hipp, J., Eggert, J., Körner, E., and König, P. (2005). Learning viewpoint invariant object representations using a temporal coherence principle. *Biological Cybernetics*, 93:79–90.

Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition*, 48:71–99.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., and Plunkett, K. (1996). *Rethinking Innateness: A Connectionist Perspective On Development*. MIT Press, Cambridge, MA.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, 3:194–200.

Franzius, M., Sprekeler, H., and Wiskott, L. (2007). Slowness and sparseness lead to place, head-direction, and spatial-view cells. *PLoS Computational Biology*, 3(8):e166.

Fukushima, K. (1980). Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202.

Fukushima, K. (1988). Neocognitron: a hierarchical neural network capable of visual pattern recognition. *Neural Networks*, 1(2):119–30.

Gaudiano, P. and Grossberg, S. (1991). Vector associative maps: unsupervised real-time error-based learning and control of movement trajectories. *Neural Networks*, 4:147–83.

Grossberg, S. (1999). How does the cerebral cortex work? learning, attention, and grouping by the laminar circuits of visual cortex. *Spatial Vision*, 12:163–87.

Harpur, G. F. (1997). *Low Entropy Coding with Unsupervised Neural Networks*. PhD thesis, Department of Engineering, University of Cambridge.

Harpur, G. F. and Prager, R. W. (1994). A fast method for activating competitive self-organising neural networks. In *Proceedings of the International Symposium on Artificial Neural Networks*, pages 412–8.

Heeger, D. J. (1991). Nonlinear model of neural responses in cat visual cortex. In Landy, M. S. and Movshon, J. A., editors, *Computational Models of Visual Processing*, pages 119–33. MIT Press, Cambridge, MA.

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9:181–97.

Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–69.

Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160:106–54.

Karni, A. (1996). The acquisition of perceptual and motor skills: a memory system in the adult human cortex. *Cognitive Brain Research*, 5:39–48.

Karni, A., Meyer, G., Jezzard, P., Adams, M. M., Turner, R., and Ungerleider, L. G. (1995). Functional MRI evidence for adult motor cortex plasticity during motor skill learning. *Nature*, 377(6545):155.

Körding, K. P. and König, P. (2001). Neurons with two sites of synaptic integration learn invariant representations. *Neural Computation*, 13(12):2823–49.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–324.

LeCun, Y., Huang, F., and Bottou, L. (2004). Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE Press.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–91.

Logothetis, N. (1998). Object vision and visual awareness. *Current Opinion in Neurobiology*, 8(4):536–44.

Lungarella, M. and Berthouze, L. (2002). On the interplay between morphological, neural and environmental dynamics. *Adaptive Behavior*, 10(3-4):223–41.

Mareschal, D., Johnson, M. H., Siros, S., Spratling, M. W., Thomas, M. S. C., and Westermann, G. (2007). *Neuroconstructivism: How the Brain Constructs Cognition*. Oxford University Press, Oxford, UK.

Marr, D. (1970). A theory for cerebral neocortex. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 176:161–234.

McClelland, J. L. and Plaut, D. C. (1993). Computational approaches to cognition: top-down approaches. *Current Opinion in Neurobiology*, 3:209–16.

Mel, B. W. (1990). *Connectionist Robot Motion Planning : A Neurally-Inspired Approach to Visually-Guided Reaching*, volume 7 of *Perspectives in Artificial Intelligence*. Academic Press.

Meng, Q. and Lee, M. H. (2007). Automated cross-modal mapping in robotic eye/hand systems using plastic radial basis function networks. *Connection Science*, 19(1):25–52.

Metta, G., Sandini, G., and Konczak, J. (1999). A developmental approach to visually-guided reaching in artificial systems. *Neural Networks*, 12:1413–27.

Mountcastle, V. B. (1998). *Perceptual Neuroscience: The Cerebral Cortex*. Harvard University Press, Cambridge, MA.

Mumford, D. (1992). On the computational architecture of the neocortex II: the role of cortico-cortical loops. *Biological Cybernetics*, 66:241–51.

Mumford, D. (1994). Neuronal architectures for pattern-theoretic problems. In Koch, C. and Davis, J. L., editors, *Large-Scale Neuronal Theories of the Brain*, pages 125–52. MIT Press, Cambridge, MA.

Oram, M. W. and Földiák, P. (1996). Learning generalisation and localisation: competition for stimulus type and receptive field. *Neurocomputing*, 11(2-4):297–321.

O'Reilly, R. C. and McClelland, J. L. (1992). The self-organization of spatially invariant representations. Technical Report PDP.CNS.92.5, Department of Psychology, Carnegie Mellon University.

Petersen, S. E., Van Mier, H., Fiez, J. A., and Raichle, M. E. (1998). The effects of practice on the functional anatomy of task performance. *Proceedings of the National Academy of Sciences USA*, 95:853–60.

Phillips, W. A. and Singer, W. (1997). In search of common foundations for cortical computation. *Behavioral and Brain Sciences*, 20(4):657–722.

Pouget, A., Deneve, S., and Duhamel, J. R. (2002). A computational perspective on the neural basis of multisensory spatial representations. *Nature Review Neuroscience*, 3:741–7.

Pouget, A. and Sejnowski, T. J. (1997). Spatial transformations in the parietal cortex using basis functions. *Journal of Cognitive Neuroscience*, 9(2):222–37.

Pouget, A. and Snyder, L. (2000). Computational approaches to sensorimotor transformations. *Nature Neuroscience*, 3(supplement):1192–8.

Quartz, S. R. (1999). The constructivist brain. *Trends in Cognitive Sciences*, 3(2):48–57.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–25.

Ritter, H., Martinetz, T., and Schulten, K. (1992). *Neural Computation and Self-Organizing Maps. An Introduction*. Addison-Wesley.

Rolls, E. T. and Milward, T. (2000). A model of invariant object recognition in the visual system: learning rules, activation functions, lateral inhibition, and information-based performance measures. *Neural Computation*, 12(11):2547–72.

Sakai, K. and Miyashita, Y. (1991). Neural organization for the long-term memory of paired associates. *Nature*, 354:152–55.

Salinas, E. and Abbott, L. F. (1995). Transfer of coded information from sensory to motor networks. *The Journal of Neuroscience*, 15:6461–74.

Schulten, K. and Zeller, M. (1996). Topology representing maps and brain function. *Nova Acta Leopoldina*, 72(294):133–57.

Serre, T., Louie, J., Riesenhuber, M., and Poggio, T. (2002). On the role of object-specific features for real-world object recognition in biological vision. In *Workshop on Biologically Motivated Computer Vision (BMCV02)*.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):411–26.

Sigala, N. and Logothetis, N. K. (2001). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, 415:318–20.

Sigman, M. and Gilbert, C. D. (2000). Learning to find a shape. *Nature Neuroscience*, 3(3):264–9.

Sinha, P. and Poggio, T. (1996). Role of learning in three-dimensional form perception. *Nature*, 384:460–3.

Siros, S., Spratling, M. W., Thomas, M. S. C., Westermann, G., Mareschal, D., and Johnson, M. H. (2008). Précis of neuroconstructivism: how the brain constructs cognition. *Behavioral and Brain Sciences*, 31(3):321–31.

Snyder, L. H., Grieve, K. L., Brotchie, P., and Andersen, R. A. (1998). Separate body- and world-referenced representations of visual space in parietal cortex. *Nature*, 394:887–91.

Spratling, M. W. (1999a). *Artificial Ontogenesis: A Connectionist Model of Development*. PhD thesis, University of Edinburgh.

Spratling, M. W. (1999b). Pre-synaptic lateral inhibition provides a better architecture for self-organising neural networks. *Network: Computation in Neural Systems*, 10(4):285–301.

Spratling, M. W. (2005). Learning viewpoint invariant perceptual representations from cluttered images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):753–61.

Spratling, M. W. (2008). Predictive coding as a model of biased competition in visual selective attention. *Vision Research*, 48(12):1391–408.

Spratling, M. W., De Meyer, K., and Kompass, R. (2009). Unsupervised learning of overlapping image components using divisive input modulation. *Computational Intelligence and Neuroscience*, 2009(381457):1–19.

Spratling, M. W. and Hayes, G. M. (1998). Learning sensory-motor cortical mappings without training. In Verleysen, M., editor, *Proceeding of the 6th European Symposium on Artificial Neural Networks (ESANN98)*, pages 339–44, Brussels, Belgium. D-Facto Publications.

Spratling, M. W. and Johnson, M. H. (2001). Dendritic inhibition enhances neural coding properties. *Cereb. Cortex*, 11(12):1144–9.

Spratling, M. W. and Johnson, M. H. (2002). Pre-integration lateral inhibition enhances unsupervised learning. *Neural Computation*, 14(9):2157–79.

Spratling, M. W. and Johnson, M. H. (2004). Neural coding strategies and mechanisms of competition. *Cognitive Systems Research*, 5(2):93–117.

Stone, J. (1998). Object recognition using spatio- temporal signatures. *Vision Research*, 38(7):947–51.

Stone, J. and Bray, A. (1995). A learning rule for extracting spatio-temporal invariances. *Network: Computation in Neural Systems*, 6(3):429–36.

Stringer, S. M. and Rolls, E. T. (2000). Position invariant recognition in the visual system with cluttered environments. *Neural Networks*, 13:305–15.

Stryker, M. P. (1991). Temporal associations. *Nature*, 354:108–9.

Templeman, J. N. and Loew, M. H. (1989). Staged assimilation: a system for detecting invariant features in temporally coherent visual stimuli. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN89)*, volume 1, pages 731–8, New York, NY. IEEE Press.

Tsotsos, J. K. (1995). Behaviourist intelligence and the scaling problem. *Artificial Intelligence*, 75:135–60.

van Rossum, M. C. W. and Renart, A. (2004). Computation with populations codes in layered networks of integrate-and-fire neurons. *Neurocomputing*, 58–60:265–70.

Wallis, G. (1996). Using spatio-temporal correlations to learn invariant object recognition. *Neural Networks*, 9(9):1513–9.

Wallis, G. (1998). Temporal order in human object recognition. *Journal of Biological Systems*, 6(3):299–313.

Wallis, G. (2002). The role of object motion in forging long-term representations of objects. *Visual Cognition*, 9:233–47.

Wallis, G. and Bülthoff, H. (1999). Learning to recognize objects. *Trends in Cognitive Sciences*, 3(1):22–31.

Wallis, G. and Bülthoff, H. (2001). Role of temporal association in establishing recognition memory. *Proceedings of the National Academy of Sciences USA*, 98(8):4800–4.

Wallis, G. and Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology*, 51(2):167–94.

Walsh, V., Ashbridge, E., and Cowey, A. (1998). Cortical plasticity in perceptual learning demonstrated by transcranial magnetic stimulation. *Neuropsychologia*, 36(4):363–7.

Weber, C., Elshaw, M., Triesch, J., and Wermter, S. (2007). Neural control of actions involving different coordinate systems. In Hackel, M., editor, *Humanoid Robots: Human-like Machines*. I-Tech Education and Publishing, Vienna, Austria.

Weber, C. and Wermter, S. (2007). A self-organizing map of sigma-pi units. *Neurocomputing*, 70(13-15):2552–60.

Weng, J., McClelland, J., Pentland, A., Sporns, O., Stockman, I., Sur, M., and Thelen, E. (2001). Autonomous mental development by robots and animals. *Science*, 291:599–600.

Westermann, G., Sirois, S., Shultz, T., and Mareschal, D. (2006). Modeling developmental cognitive neuroscience. *Trends in Cognitive Sciences*, 10:227–33.

Wiskott, L. and Sejnowski, T. J. (2002). Slow feature analysis: unsupervised learning of invariances. *Neural Computation*, 14(4):715–70.