Network: Computation in Neural Systems
September 2014; 25(3): 116–136

informa
healthcare

Original Article

Self-organization of head-centered visual responses under ecological training conditions

BEDEHO M. W. MENDER & SIMON M. STRINGER

Department of Experimental Psychology, University of Oxford, UK

(Received 14 February 2014; revised 9 April 2014; accepted 23 April 2014)

Abstract
We have studied the development of head-centered visual responses in an unsupervised self-organizing neural network model which was trained under ecological training conditions. Four independent spatio-temporal characteristics of the training stimuli were explored to investigate the feasibility of the self-organization under more ecological conditions. First, the number of head-centered visual training locations was varied over a broad range. Model performance improved as the number of training locations approached the continuous sampling of head-centered space. Second, the model depended on periods of time where visual targets remained stationary in head-centered space while it performed saccades around the scene, and the severity of this constraint was explored by introducing increasing levels of random eye movement and stimulus dynamics. Model performance was robust over a range of randomization. Third, the model was trained on visual scenes where multiple simultaneous targets where always visible. Model self-organization was successful, despite never being exposed to a visual target in isolation. Fourth, the duration of fixations during training were made stochastic. With suitable changes to the learning rule, it self-organized successfully. These findings suggest that the fundamental learning mechanism upon which the model rests is robust to the many forms of stimulus variability under ecological training conditions.

Keywords: *Head-Centered Response, Neural Network, Self-Organization, Ecological Validity*

Introduction

The ability to process sensory input, make a decision about the appropriate course of action, and execute the corresponding motor commands, is critical for the survival of an animal. The critical step of translating sensory input encoded in a

Correspondence: Bedeho M. W. Mender, Department of Experimental Psychology, University of Oxford, UK. E-mail: bedeho.mender@psy.ox.ac.uk

ISSN 0954-898X print/ISSN 1361-6536 online/04/030116–136 © 2014 Informa UK Ltd.
DOI: 10.3109/0954898X.2014.918671

modality specific neural representation, for example about the location of targets or threats, into the neural representation of the behaviourally relevant motor effector typically involves a coordinate transformation. In primates, vision is a key sensory input used for guiding survival behaviours like evasion, feeding and mating. Visual signals gathered at the retina are initially encoded in an eye-centered reference frame, meaning that receptive fields are anchored to the point of fixation. However, motor commands like reaching or gaze adjustment require supra-retinal reference frames. A substantial body of theoretical work has been devoted to understanding how such coordinate transformations may occur in a number of parietal areas which contain the relevant sensory and motor signals (Zipser and Andersen 1988; Mazzoni et al. 1991; Pouget and Sejnowski 1997; Xing and Andersen 2000). However, the overwhelming majority of this work was based on supervised error correction algorithms which did not provide any plausible hypothesis for how such coordinate transformation circuits may develop in cortex.

To address this, we have previously proposed and modelled a biologically plausible self-organization hypothesis for how the transformation of visual signals from an eye-centered to a head-centered reference frame may occur in primate parietal areas (Mender and Stringer 2013). Because a primate adjusts its gaze more frequently by moving its eyes rather than its head (Freedman and Sparks 1997; Einhäuser et al. 2007), there will be periods of time during which visual targets within a scene will remain stationary with respect to the head while shifting on the retina. This results in temporal sequences of visual inputs with the targets in fixed locations in head-centered space but occurring in different eye-centered locations. In our model, these visual input patterns are encoded by a population of retinotopic neurons with eye position gain modulation as have been found, for example, in areas PO (V6) (Galletti et al. 1995), 7a and LIP (Andersen et al. 1990). These input neurons send feedforward connections to a layer of output neurons. The feedforward connections from the input layer to output layer are modified during visual training by a *trace learning* rule (Foldiak 1991), which incorporates a memory trace of recent neuronal activity. The effect of the trace learning rule is to encourage individual output neurons to learn to respond to subsets of visual input patterns that tend to occur close together in time. During training, the model performs sequences of saccades around a number of visual scenes with targets in fixed locations while the head remains stationary. In this case, trace learning encourages individual output neurons to learn to respond to a subset of temporally proximal input patterns corresponding to the same head-centered target location. In this way, individual output neurons learn to respond to input patterns corresponding to a visual target in a specific head-centered location, thus endowing these neurons with head-centered responses.

The basic approach of temporal binding by some form of trace learning has also been explored in an earlier study by (Spratling 2009). However, this model did not explore the impact of the full range of ecological training conditions investigated here, which is a critical consideration given that the model depends on these assumptions for proper self-organization.

Our previous study (Mender and Stringer 2013) provided only a basic demonstration of the underlying computational hypothesis under a number of highly idealized and rather unrealistic training conditions. In particular, the framework was not tested under more natural conditions of visually-guided training. This is a key issue because the self-organisation of the network architecture

depends critically on the statistics of natural eye and head movements, as well as the statistics of visual scenes. Thus, testing model self-organization under more realistic conditions is an essential next step and the subject of the current paper.

For example, in our earlier work (Mender and Stringer 2013), the model was only exposed to eight different head-centered locations during training, while under natural conditions the visual targets would be located across the full continuum of head-centered space. Whether or not the model can continue to develop head-centered representations when trained on a continuum of such locations is non-trivial because in this situation an invariance learning mechanism known as continuous transformation (CT) learning may begin to degrade the network performance by causing output neurons to respond over very large regions of the head-centered space (Stringer et al. 2006). Such invariance learning may even eliminate head-centered responses altogether. An important question, then, is whether output neurons remain head-centered, and if so, what happens to the size of the receptive fields as the number of training locations approaches a continuum.

Another critical issue in our earlier study (Mender and Stringer 2013) was that the model was only exposed to training periods in which the head and visual targets remained perfectly stationary during a sequence of saccadic eye movements. However, under natural conditions such sequences of saccadic eye movements would be frequently interrupted by additional periods in which the head or visual targets are also moving. The basic self-organization hypothesis sets strict conditions on the statistics of how the eyes and head move naturally. Specifically, the eyes must move rapidly while the head and visual objects in the world remain fixed. So will head-centered representations still develop if for much of the time the statistics of eye and head movement depart significantly from this ideal? The current study begins to investigate this question.

An even more fundamental limitation of our previous work (Mender and Stringer 2013) was that the model was only exposed to a single visual target in the scene during training at any given time, while natural scenes always include multiple targets. How could output neurons learn to respond to single head-centered locations if the model is always exposed to stimuli in multiple locations during visually-guided learning? This is similar to the problem of how neurons in the primate ventral visual pathway learn to represent individual objects when trained on natural scenes with multiple objects present (Stringer and Rolls 2000). Some authors have suggested that attentional mechanisms are needed to highlight one stimulus location at a time during learning (Rolls and Deco 2002). Below, we show that our model can successfully self-organise head-centered output neurons without such an attentional mechanism.

Lastly, in our earlier work (Mender and Stringer 2013), the model always performed uniform 300 ms eye fixations interleaved with saccades, while under natural conditions there would be much greater variability in fixation durations. Interestingly, this turns out to be a potential challenge for the basic self-organisational hypothesis. In the simulations reported below it was found that the model failed to develop head-centered output neurons unless some additional architectural modification was introduced to the model. Again, the reason for this is that the self-organization depends critically on the statistics of eye and head movement. For example, if the fixations last too long, then trace learning may fail to bind across successive fixations. Below we show one biologically plausible way in which this problem may be ameliorated in the brain.

To begin to address these issues, this paper investigates whether the model can cope with a greater level of variability within each of these ecologically important dimensions of the visual training stimuli. Below we show that the model is indeed able to successfully develop head-centered output representations when trained under these more ecologically realistic training conditions. The robust performance of the model lends support to the proposal that these learning mechanisms are operating in the primate cortex.

Materials and methods

Network architecture

The architecture of the neural network model and methods for analysing its performance are similar to our earlier study (Mender and Stringer 2013). The network consisted of an input population that sent feedforward projections to an output population as shown in Figure 1. The input neurons encoded the retinal locations of visual stimuli, where the responses were modulated by the position of the eye in its orbit. The retinal location and eye position spaces covered $[-90^\circ, 90^\circ]$ and $[-30^\circ, 30^\circ]$, respectively. The N output neurons were a competitive layer (Rolls and Treves 1998). Each neuron in the output layer received connections from a randomly chosen subset of φ percent of the input population. There was no topagraphic arrangement of neurons or connections within the model.

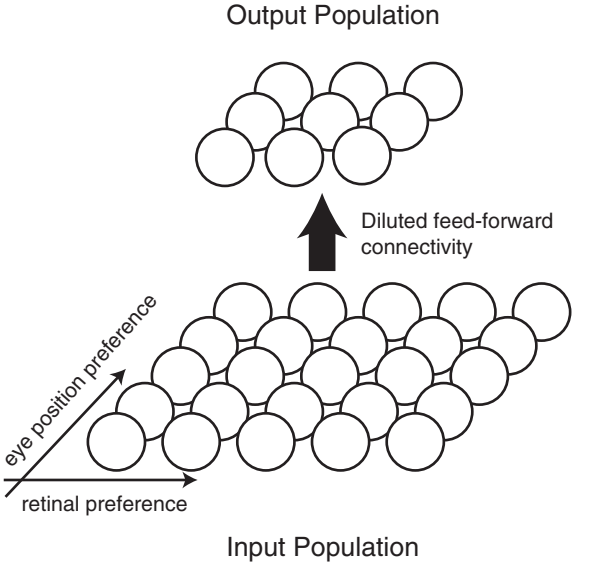

Figure 1. Architecture of 2-layer neural network model, where the population of input neurons projects to the competitive output population. Neurons in the input population have a unique combined eye position and retinal preference, but the population has no topographic spatial organization. Each output neuron is connected to a subset of neurons in the input population.

Training the network

At the start of training, the synaptic connections were set to random initial values. The network was then trained on a series of visual scenes containing one or more visual targets, during which the model performed eye and head movements. In each simulation, the visual targets were located in M evenly spaced head-centered locations within $[-63^\circ, 63^\circ]$, which ensured that the visual targets always remained within the field of view.

Each training epoch was comprised of a number of periods, where in each period there was a fixed unique subset of k head-centered locations occupied by visual targets. During such a period, the locations of the k visual targets with respect to the head remained fixed while the eyes saccaded through a randomised sequence of P different eye positions. In most experiments the model was trained on only one visual target at a time, hence there were M periods within each training epoch, each corresponding to a visual target being located in one of the M head-centered locations. In other experiments the model was trained on two or more targets ($k > 1$) presented simultaneously within each period. In this case, since the model was trained on k visual objects presented in all possible subsets of the M head-centered locations, there were $\binom{M}{k}$ such periods within each training epoch. During each period in which k visual targets were presented in a fixed combination of head-centered locations, the model fixated P uniformly sampled eye positions in $[-24^\circ, 24^\circ]$. The duration of each fixation was usually set to 300 ms, although in some experiments this was varied. The model saccaded between successive eye positions at 400°/s.

The above training protocol obeyed the key assumption that the eyes move more frequently than the head. However, in some simulations reported below we explored the effect of relaxing this condition for a portion of the training time.

Testing the network

After the synaptic connections had been set up during training, the model was tested to see whether the output neurons had developed head-centered response properties. In particular, we tested the reference frame of the neuronal response, receptive field size and receptive field location. It was important, however, to test the ability of the model to generalise to new combinations of target location and eye position, which the model had not been trained on. To do this, the responses of the output neurons were recorded as the model fixated in $E = 4$ eye positions -18°, -6°, 6° and 18°, during which a single visual target was placed in each of $T = 80$ head-centered target locations within $[-79^\circ, 79^\circ]$ in increments of 2°.

Input population

The input neurons encoded the retinal locations of visual targets, where the neuronal responses were modulated by the position of the eye in its orbit. Neurons with these firing properties have been reported in brain areas such as PO (Galletti et al. 1995),

7a and LIP (Andersen et al. 1990). Specifically, the firing rate of each input neuron i was given by

$$v_i^I = \exp\left(-\frac{||e-\beta_i||^2}{2\rho^2}\right) \times \sum_{r\in R} \exp\left(-\frac{||r-\alpha_i||^2}{2\sigma^2}\right) \quad (1)$$

where e is the current eye position, and r denotes the retinal locations of a number of visual targets. It can be seen that the neuronal response depends on a product of two terms. The first term represents an eye position gain field, where β_i is the the neuron's preferred eye position and ρ is the standard deviation of the corresponding Gaussian tuning curve. This form of peaked eye position gain field has been found in area PO (Galletti et al. 1995). The second term reflects the neuron's response to the presence of R visual targets in the scene, where α_i is the neuron's preferred retinal location and σ is the standard deviation of this Gaussian tuning curve. The population of input neurons covered the two dimensional space of all integer combinations of eye position and retinal target location.

Output population

Each output neuron i had three variables defined: a trace $q_i(t)$, an activation $h_i(t)$, and a firing rate $v_i(t)$ (Dayan and Abbott 2001).

The activation was set according to

$$\tau_h \frac{\mathrm{d}h_i}{\mathrm{d}t} = -h_i + \sum_j w_{ij} v_j^I \quad (2)$$

where τ_h was the time constant, and w_{ij} was the synaptic weight from input neuron j to output neuron i.

The firing rate was set according to the sigmoid function

$$v_i = \frac{1}{1+\exp(2\varphi(h_i - p_\pi - \theta)} \quad (3)$$

with slope φ and threshold θ.

The parameter p_π controlled competition between output neurons by setting the proportion of neurons that remained active. Specifically, p_π was set to the activation value at the πth percentile point of the distribution of neuronal activations. So, for example, setting π to 90 ensured that approximately 10 per cent of the output neurons remained active.

Trace learning

Trace learning rules encourage postsynaptic neurons to respond to clusters of input patterns that tend to occur close together in time. To achieve this, trace learning rules incorporate a trace of recent neuronal activity. In our model, the trace value q_i (t) of output neuron i was computed according to

$$\tau_q \frac{\mathrm{d}q_i}{\mathrm{d}t} = -q_i + v_i \quad (4)$$

where τ_q was a time constant.

During training, the strength of the synaptic weight w_{ij} (t) from input neuron j to output neuron i was modified according to the trace learning rule

$$\frac{dw_{ij}}{dt} = Qq_i v_j^I \tag{5}$$

where Q was the learning rate, v_j^I was the firing rate of input neuron j, and q_i was the trace value of output neuron i.

To prevent unbounded growth of the synaptic weights during training, after each weight update the weight vectors of all output neurons were renormalised, as is typical in competitive neural networks (Rolls and Treves 1998).

The trace learning rule encourages output neurons to bind together clusters of input patterns that tend to occur in temporal proximity. If the agent moves its eyes more frequently than its head, as per our hypothesis, then input patterns corresponding to the same fixed head centered target location will tend to be clustered together in time. In this case, a trace learning rule will encourage output neurons to learn to respond to visual targets in a head centered reference frame.

Numerical simulation of the differential model equations

The differential Equations 2, 4 and 5 were solved using a Forward-Euler finite difference scheme, where the numerical time step Δt was set to one tenth of τ_h. The visual and eye-position signals required as inputs to the model were simulated dynamically and sampled at 1 kHz. Where necessary, linear interpolation was then used to compute the numerical inputs to the discretized Forward Euler equations.

Analysis of model behaviour after training

After training the network, we tested whether the output neurons had learned to respond to the presence of visual targets in a head-centered frame of reference. The methods used for analysing the performance of the model were similar to those used in our earlier study (Mender and Stringer 2013). We recorded the responses of each output neuron over every combination of the E eye positions and T head-centered target locations. For each output neuron we computed the matrix $\mathbf{R}$, where $\mathbf{R}[i, j]$ denotes the neuronal response for the ith eye position e_i and jth head-centered target location t_j. The vector $(\mathbf{R}[i, 1], \ldots, \mathbf{R}[i, T])$ is referred to as the response vector at the ith eye position.

Reference frames. The primary question was whether the population of output neurons had learned to respond in either an eye-centered reference frame of head-centered reference frame after training. To quantify this, two metrics were computed for each output neuron. These metrics reflected how compatible the responses of the neuron were with either a head-centered or eye-centered reference frame, respectively.

The head-centeredness metric, Π, reflected the stability of neuronal responses to head-centered target locations across the E different eye positions. Details of how this was computed are presented in our previous study (Mender and Stringer 2013). The metric was bounded in the interval $[-1,1]$, where a value of 1 indicated a

perfect head-centered response. This metric was referred to as the *head-centeredness* of the output neuron.

A similar approach was taken to computing the eye-centeredness metric, Ω, which reflected the stability of the neuronal responses to retinal target locations across the E different eye positions. How this was computed is described in our previous study (Mender and Stringer 2013). The metric was similarly bounded in $[-1,1]$, where a value of 1 indicated a perfect eye-centered response. This metric was referred to as the *eye-centeredness* of the output neuron.

The two metrics, Π and Ω, were finally combined into a single unidimensional measure known as the receptive field index (RFI) as follows

$$\mathrm{RFI} = \begin{cases} \Pi - \Omega & \text{if } 0 \leq \Pi \leq 1 \text{ and } 0 \leq \Omega \leq 1 \\ \Pi & \text{if } 0 \leq \Pi \leq 1 \text{ and } -1 \leq \Omega \leq 0 \\ -\Omega & \text{if } -1 \leq \Pi \leq 1 \text{ and } 0 \leq \Omega \leq 1 \\ 0 & \text{if } -1 \leq \Pi \leq 1 \text{ and } -1 \leq \Omega \leq 1 \end{cases} \tag{6}$$

The RFI was also bounded between -1 and 1. If an output neuron had a strictly positive RFI, then it was classified as head-centered. Alternatively, if the RFI was strictly negative then the output neuron was classified as eye-centered. Otherwise, if the RFI was zero, then the neuron remained unclassified.

Another performance metric was the *head-centeredness rate*, which was the proportion of output neurons that were classified as head-centered with RFI >0.

Head-centered receptive field location. The head-centered receptive field location of an output neuron was determined in the following way. For each eye position e_i for $i = 1, \ldots, E$, the centre of mass of the head-centered response vector ($\mathbf{R}[i, 1], \ldots, \mathbf{R}[i, \mathrm{T}]$) was computed across the T head-centered target locations. Then the head-centered receptive field location was computed as the average of these centres of mass over all E eye positions as follows

$$\frac{1}{E}\sum_{i=1}^{E}\frac{\sum_{j=1}^{T} t_j \mathbf{R}[i,j]}{\sum_{j=1}^{T}\mathbf{R}[i,j]} \tag{7}$$

Coverage. If the model is performing well, then the receptive field locations of head-centered output neurons should ideally be evenly distributed over the M head-centered training locations $g_1, \ldots, g_M$. In order to assess this, the first step was to compute the head-centered receptive field location for each head-centered output neuron as described above. Then each such neuron was assigned to the nearest of the M head-centered training locations. Let p_i denote the fraction of head-centered neurons assigned to training location g_i. Then the *coverage* of the model was defined as the normalized entropy of this distribution

$$-\frac{1}{\log_2 M}\sum_{i=1}^{M} p_i \log_2 p_i \tag{8}$$

If there was a perfectly uniform distribution, then this measure would give a maximal coverage of 1. However, if there was some $p_i = 0$, where the ith head-centered training location was not represented by any output neurons, then the coverage was not defined and there was said to be no coverage.

Receptive field size. Another important diagnostic was the size of the head-centered receptive fields of output neurons after training. For this analysis, a neuron was considered responsive whenever its firing rate was above 50% of the neuron's maximal response across all eye positions and head-centered target locations. For each eye position, the size of the receptive field was computed by adding up all regions of the head-centered space in which the neuron responded. The final receptive field size of the given neuron was then computed as the average receptive field size across the different eye positions. Further details of this calculation are provided in our earlier study (Mender and Stringer 2013).

Results

Number of target locations

In natural visual scenes, visual targets can be seen in any location with respect to the head. Therefore, it was important to verify that the self-organizing model could operate under these more ecological conditions. In this experiment it was investigated how varying the number of visual target locations in head centered space during training, M, would influence self-organization in the model. In particular, the asymptotic performance of the model was studied as the number of head centered target locations was increased towards an effective continuum, whereby visual targets may be seen anywhere with respect to the head.

Each simulation was performed with a fixed value of M. For each simulation, a stimulus set was created as described in section 2. Each training epoch was divided into M periods, each of which corresponded to one of the head centered target locations. During each such training period, the location of the visual target remained fixed in the corresponding head centered location while the model saccaded through a sequence of $P = 15$ eye positions. Three example stimulus sets are shown in Figure 2 for values of M of 2, 4 and 12. Table I gives the full parameter set of the experiment.

The impact of varying the number of head centered target locations during training on the performance of the model was inspected by plotting key summary statistics as a function of M in Figure 3. The head-centeredness rate and the average head-centeredness increased approximately monotonically with increasing M, and were always above their corresponding untrained network values of $\sim$25% and $\sim$0.17 respectively. The coverage did not drop below $\sim$0.88 when $M \geq 7$ in the trained model, while there was no coverage for the untrained model. The average head centered receptive field size decreased steadily as M increased, and was always well below the untrained average of $\sim 69^\circ$. In summary, these results showed that model performance actually improved as the number of head centered training locations was increased during training.

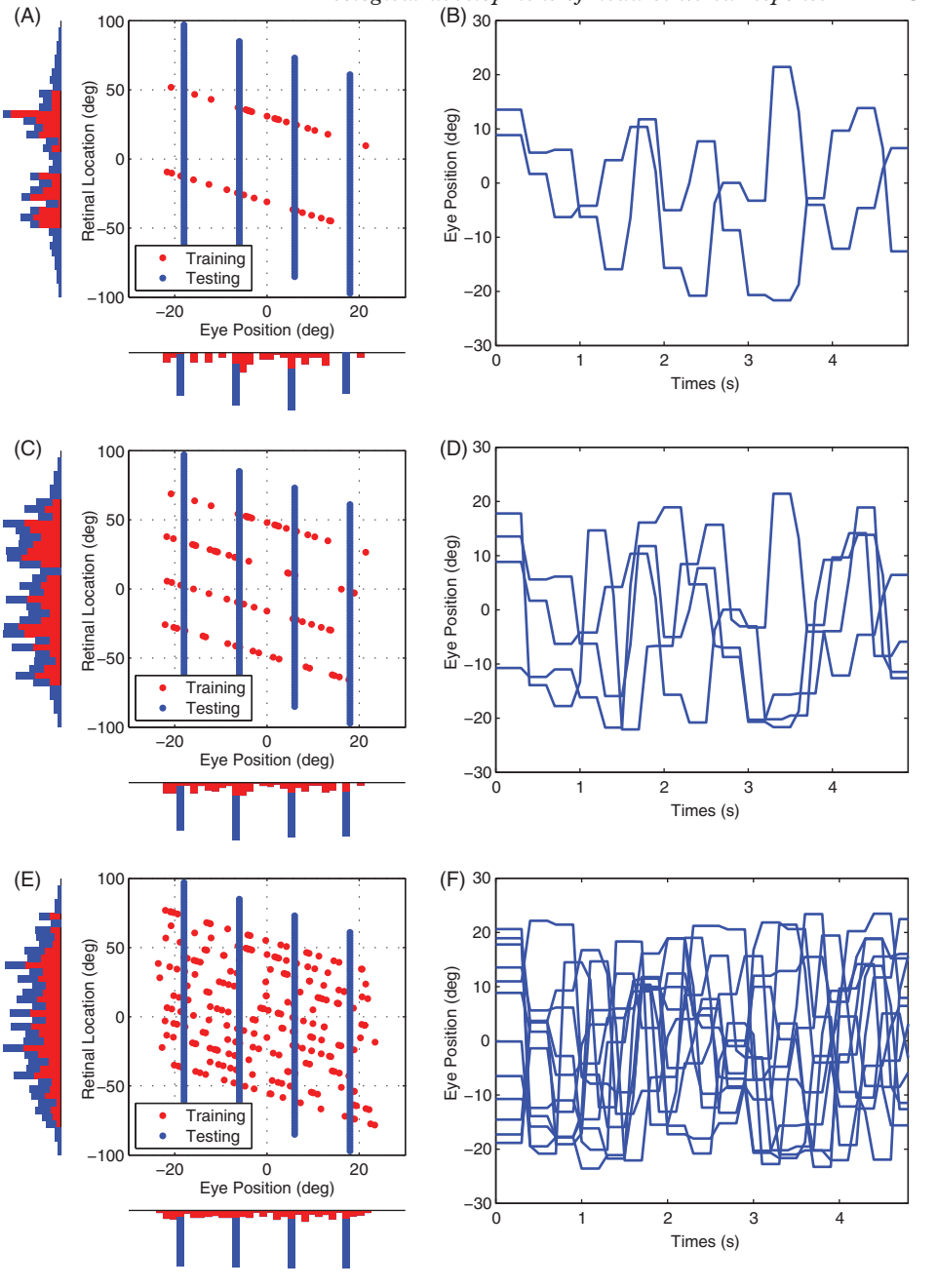

Figure 2. Simulated movements of the eyes and head-centered locations of visual targets during training and testing, and each row shows the stimuli for a given value of *M*, that is 2, 4 and 12 from the top respectively. The left hand column (A,C,E) shows scatter plots in which each point corresponds to a single fixation during either training (red points forming negative slope line) or testing (blue points forming four vertical lines). The fixation points are plotted as a function of the eye position (abscissa) and the retinal location of the visual target (ordinate). Each of the diagonal lines of red points corresponds to a period during training when the visual target was fixed in one of the head-centered target locations while the eyes moved. The vertical lines of blue points correspond to the four eye positions in which the network was tested. The right hand column (B,D,F) shows plots showing how the eye position is shifted through time in a randomised manner during training. Each trace corresponds to a different period during which the visual target is maintained in a fixed head centered location.

Table I. Parameters of experiment varying the number of target locations during training (M).

Parameter	Symbol	Value
Number of target locations	M	1,...,30
Fixation sequence length	P	15
Number of training epochs	–	20
Width of eye position tuning curve	ρ	6°
Width of retinal tuning curve	σ	6°
Output neuron population size	N	900
Input neuron population size		12261
Trace time constant	τ_q	400 ms
Activation time constant	τ_h	100ms
Activation function slope	φ	4.5
Activation function threshold	θ	0.4
Sparseness percentile	π	80%
Learning rate	Q	0.05
Synaptic connectivity	Φ	5%

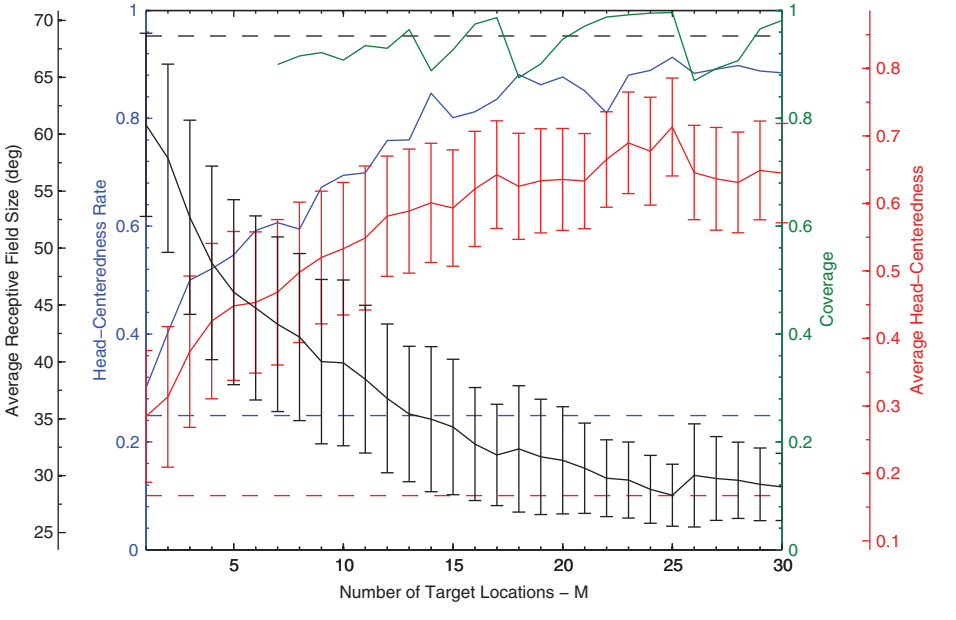

Figure 3. The plot shows four key population metrics as a function of M. The average receptive field size curve (black declining curve with error bars) shows the average size of the head centered receptive field among head-centered neurons, and the error bars represent the standard deviations. The head-centeredness rate (blue increasing curve without error bars) shows the fraction of output neurons that were head-centered. The coverage curve (green curve appearing at $M = 7$) shows the coverage of the head-centered training locations by the output neuron population, where missing data points before epoch $M = 7$ were due to at least one of the eight head-centered training locations not being represented by the output cells. The average head-centeredness curve (red increasing curve with error bars) shows the average head-centeredness value among all head-centered neurons, and the error bars were the standard deviations. The dashed lines show the corresponding quantity in the untrained model, that is average head-centeredness, head-centeredness rate and average receptive field size respectively, from bottom to top in plot.

Variation in statistics of stimulus dynamics

The basic hypothesis for how head-centered output representations may develop assumes that during natural self-motion, there are periods of time when the eyes are moving in the head while the head remains stationary with respect to the visual environment and visual objects also remain stationary within the environment. In the simulations reported so far, the statistics of the stimulus dynamics have conformed to this assumption. However, in reality there will be times when the stimulus dynamics radically depart from this assumption. For example, sometimes the eyes might track a moving object while the head remains stationary, or the head may move while the eyes fixate a stationary object. Both of these alternative stimulus dynamics do not result in visual objects remaining in a fixed head-centered location for a continuous period while the object shifts on the retina, which is required for trace learning to form head-centered output representations. However, the hypothesis does not actually require the stimulus dynamics to conform to this assumption *all* of the time. Indeed, it was conjectured that as long as there were some periods of time when the visual objects remain in a fixed head-centered location while the eyes move, which may be interspersed with periods governed by alternative stimulus dynamics, then this would still permit head-centered output representations to develop. This key property is important to verify in order to support the biological and ecological validity of the model.

The following experiment investigated how altering the stimulus dynamics by including additional periods of randomised visual target and eye movements would influence the self-organization of the model. In previous experiments, the input stimulus was structured into regular training periods corresponding to sequences of eye movement while the stimulus was kept stationary with respect to the head. In our previous work (Mender and Stringer 2013) it was established that by reducing the length of these periods, effectively decreasing the relative frequency of eye versus head movements, the model performed worse in terms of the number of head-centered output neurons developed through learning. In the following experiment, the regular training periods during which the eyes move while the visual target remains fixed with respect to the head were interleaved with new periods consisting of completely randomised eye and stimulus movement sequences. It was investigated how the self-organization was influenced by varying the length of these new periods of randomised stimulus dynamics. The expectation was that increasing the length of these randomised stimulus sequences would eventually dominate the effects of the regular training periods promoting the development of head-centered neurons. How tolerant the model was to increasing the proportion of randomised stimulus dynamics during training would inform whether the model could self-organise head-centered output representations under more natural training conditions.

An epoch of training had $M = 8$ regular training periods where a single visual target was located in one of the eight head-centered locations $-56°$, $-40°$, $-24°$, $-8°$,$8°$,$24°$,$40°$ or $56°$ while a sequence of $P = 30$ fixations were performed. Additionally, the M regular training periods were interleaved by randomised training periods during which there was a fixed number, K, of random movements of both the eyes and the location of the stimuli in head-centered space. Separate simulations were performed for $K = 10, 15, \ldots, 75$ to explore the effects of gradually increasing the proportion of training that was governed by randomised stimulus

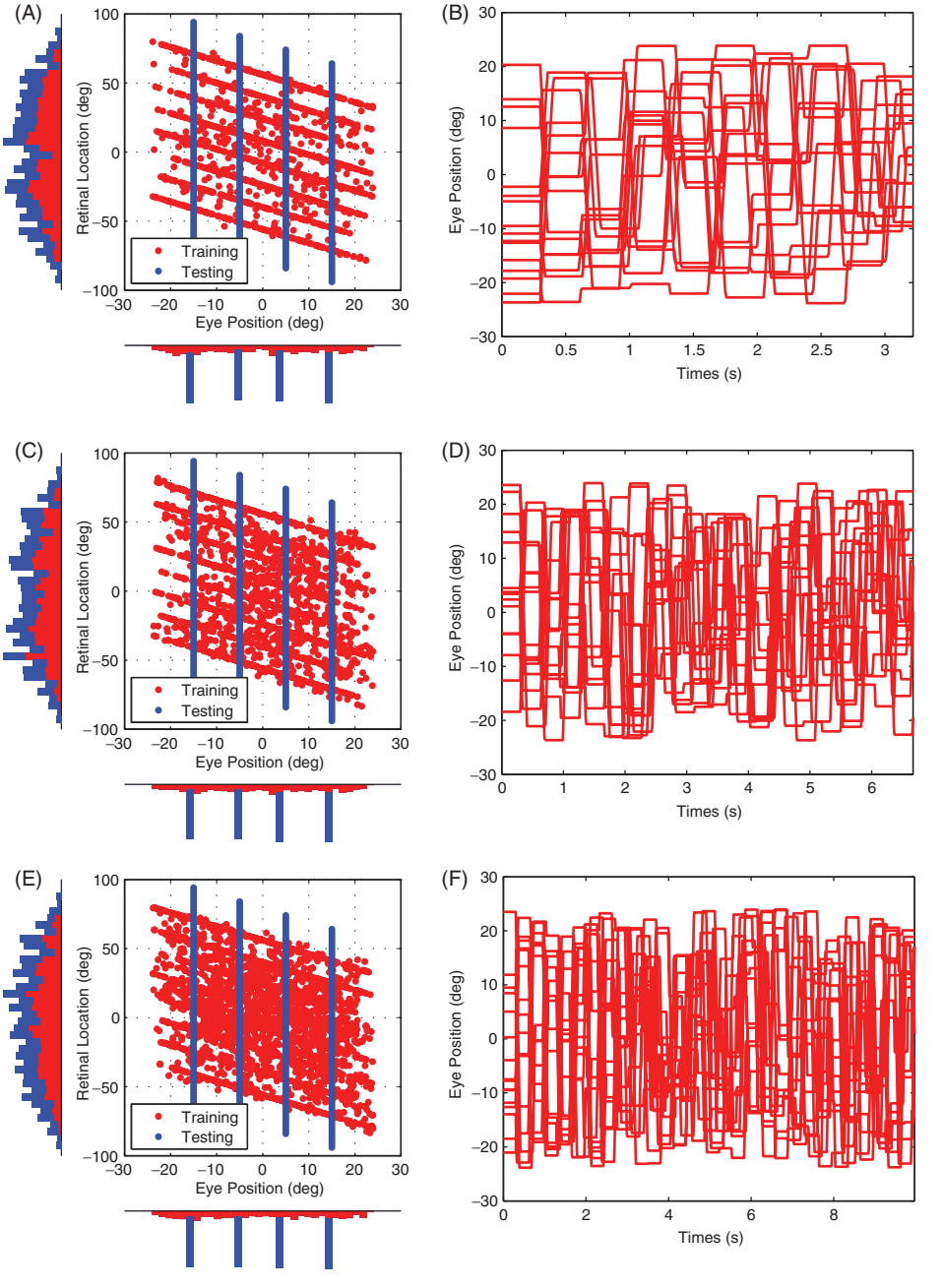

Figure 4. Movements of the eyes and locations of visual targets during simulations with additional training periods of randomised movements of the visual target and eyes. Each row shows the input stimuli for a given value of K, that is 10, 20 and 40 from the top respectively. The plots show the input stimuli during both periods of regular movements and periods of randomised movements. The left column (A,C,E) shows the fixation positions in the testing and training set. The fixations off the main diagonals in the left column plots represent the fixations that occur during the training periods with randomised movements of the visual target and eyes. The right column (B,D,F) shows the eye movement dynamics during each period within a single training epoch. The eye movement dynamics are shown only for the duration of the shortest period within each plot. That is, for $K = 10$ and $K = 20$ this corresponds to the periods of randomised movements, while for $K = 40$ this corresponds to the periods of regular movements for which $P = 30$.

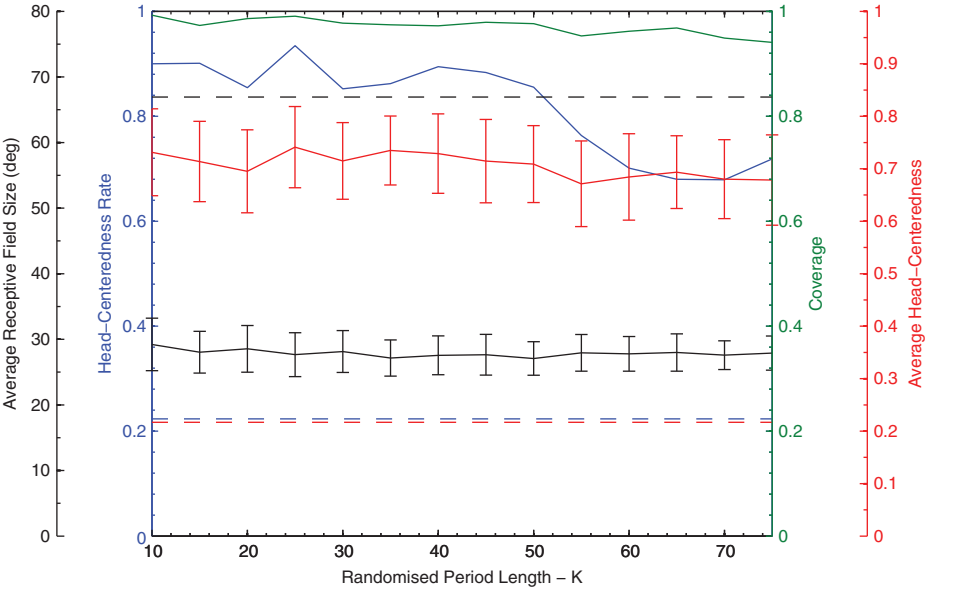

Figure 5. Population summary statistics of the response properties of the output neurons from experiments with additional training periods with randomised movements of the visual target and eyes. Results are plotted for simulations with different lengths of the randomised training periods *K*.

dynamics. Figure 4 shows the training and testing stimuli for three different values of *K*, namely 10, 20 and 40. The simulation parameters were otherwise the same as the previous experiment.

The effects of varying the length of the randomised training periods on the performance characteristics of the model was inspected by plotting key summary statistics as a function of *K* in Figure 5. The head-centeredness rate remained above 67% across the explored range of *K*. While it did eventually decline with increasing *K*, it was still well above the head-centeredness rate of ~22% in the untrained model. There was coverage across all explored values of *K*, and the coverage remained close to maximal, i.e. greater than ~0.94, for all *K*. Among head-centered neurons the average head-centeredness remained very stable and never went below 0 67. Indeed, this measure showed minimal dependence on *K*. It was also well above the average head-centeredness among head-centered neurons in the untrained model which was ~0.22. The average receptive field size among head-centered neurons also remained very stable and close to 30°, which was well below the untrained network value of ~67°. In summary, this showed that model performance was robust to the introduction of a significant level of randomised eye- and head-movements during training.

Multiple simultaneously visible targets during training

In this experiment it was investigated whether the model could self-organize to develop neurons with head-centered receptive fields despite always being exposed

to multiple simultaneously appearing visual targets during training. This was an important issue to explore because the primate visual system will typically be exposed to multiple objects within a visual scene at any one time. So it is important to verify that the model can self-organize successfully under these conditions.

Stimuli were presented during training in *two* target locations at a time while the eyes moved through a sequence of saccades. Showing stimuli in all possible pairs of target locations was intended to break the statistical coupling between different target locations. This effective statistical decoupling between different target locations can be exploited by the architecture and operation of a biologically plausible competitive neural network, which will be forced to learn to represent the individual target locations rather than the pairs of target locations that are actually presented during training. That is, the competitive output layer is forced to develop output neurons that only respond to one target location. A number of papers have previously explored how this general phenomenon of statistical decoupling between multiple visual objects can allow a competitive neural network model of the primate visual system to develop separate representations of the individual objects that have been shown together in different combinations (e.g. pairs or triples) during training (Stringer et al. 2007, 2008). In particular, (Stringer et al. 2007) showed that this mechanism of statistical decoupling in a competitive neural network architecture can be successfully combined with trace learning to simultaneously perform temporal binding of input patterns that occur in temporal proximity.

In the following simulation there were eight head-centered target locations that the visual stimuli could occupy during training. During training, stimuli were presented in two of these locations at a time while the eyes shifted through a sequence of saccades. Therefore, each epoch of training consisted of $\binom{8}{2} = 28$ periods, where the network was exposed to a unique pair of target locations for each such period. In particular, none of the target locations were ever presented singularly to the network during training. Unlike previous experiments, the sequence of eye movements was identical across periods, which was found to permit statistical decoupling with fewer fixations.

Figure 6 shows the population results from testing the model before and after training, and population statistics are given in Table II.

The head-centeredness rate increased from ~35% to ~97% after training (Figure 6A), and this was also significantly higher than the head-centeredness rate of ~69% found when training only with a single visible target. Among the head-centered neurons, the average head-centeredness increased from 0.08 to 0.59 with training. The receptive fields were clustered around the eight training locations (Figure 6C). Neurons did not have multi-modal receptive field peaks corresponding to multiple target locations, but rather neurons were tuned to single training locations as desired. This was also reflected in a very even distribution of head-centered receptive fields among the eight locations, which gave a coverage of ~0.99. In summary this showed that the model could successful develop head-centered output neurons with single peaked receptive fields in head-centered space after being exposed only to *multiple* (i.e. pairs of) visual targets appearing simultaneously during training.

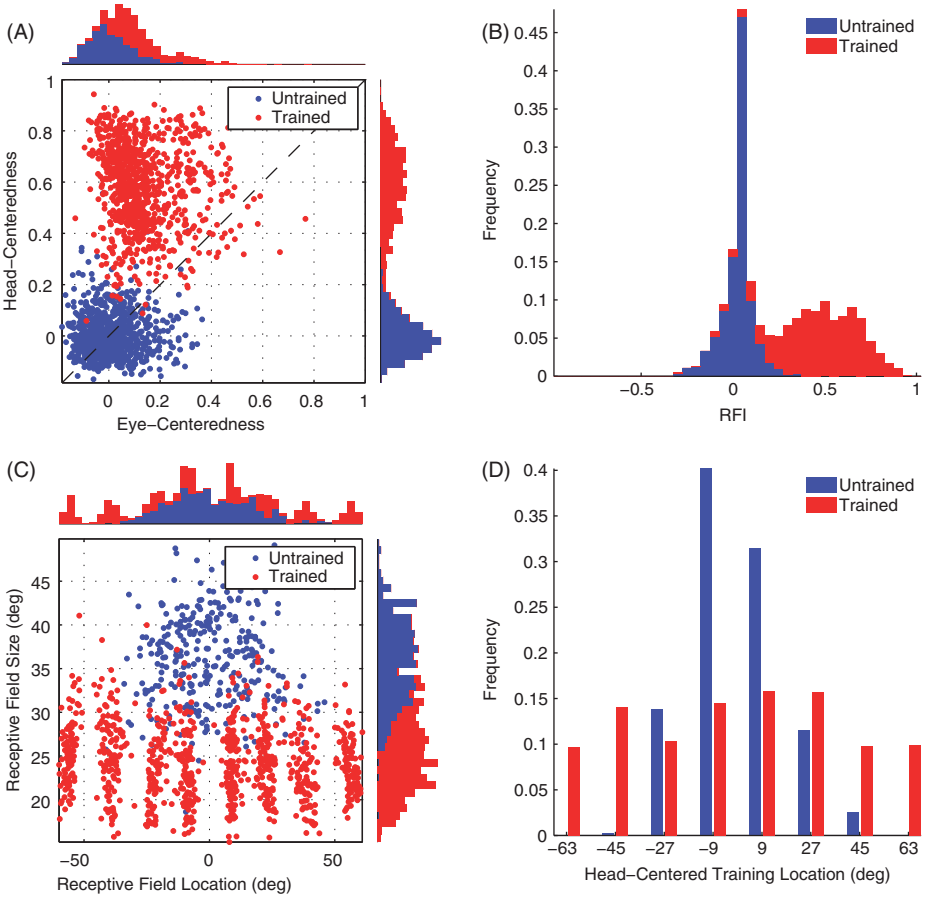

Figure 6. Population analyses of receptive field properties of output neurons in the untrained and trained model from experiment with multiple simultaneously visible targets during training. (A) Scatter plot shows the reference frame response characteristics of all neurons in the output layer, where each neuron is plotted as a point corresponding to that neuron's particular combination of head-centeredness and eye-centeredness. Data points for the untrained model are plotted in blue, while results for the trained model are shown in red. (B) Distributions for receptive field index values in the two models. (C) Scatter plot showing the combination of head centered receptive field size and head-centered receptive field location of all head-centered output neurons for the two models. (D) Histograms showing the distribution of the numbers of output neurons that responded preferentially to each of the head-centered locations which were observed during training.

Variability in fixation durations

In this experiment it was investigated how variability of fixation duration within the same model simulation would influence the self-organization of the model. In the previous experiments the fixation duration was fixed, however, under natural conditions there would be substantial variability in the duration of individual fixations. Therefore, it was important to establish whether the model could still self-organize successfully under these conditions. The training dynamics were exactly as before, except that all fixation durations were independently sampled from

Table II. Population summary statistics of the response properties of output neurons in three different conditions. Results for the untrained model are shown in the left two columns. Results for the model trained with a single visible target during training are shown in the middle two columns. Results for the model trained with multiple simultaneously visible targets during training are shown in the rightmost two columns. For each type of model, results are presented in two subcolumns: statistical measures computed over all output neurons are shown in the left subcolumn, while measures computed over neurons with a receptive field index greater than zero indicating head-centered responses are shown in the right subcolumn. Each row corresponds to a different performance metric: head-centeredness, eye-centeredness, receptive field index, head-centered receptive field location, and head-centered receptive field size. Each cell of the table shows the mean and standard deviation (in parentheses) of the performance metric over the relevant population of output neurons.

	Experiment 3.3					
	Untrained		Trained (single target)		Trained (multiple targets)	
	All	RFI >0 (−35%)	All	RFI >0 (−69%)	All	RFI >0 (−97%)
Head-centeredness	0.00 (0.08)	0.08 (0.07)	0.58 (0.19)	0.63 (0.16)	0.58 (0.16)	0.59 (0.15)
Eye-centeredness	0.01 (0.09)	−0.03 (0.06)	0.36 (0.24)	0.25 (0.15)	0.13 (0.12)	0.12 (0.11)
RFI	−0.01 (0.09)	0.07 (0.07)	0.22 (0.31)	0.38 (0.20)	0.45 (0.21)	0.47 (0.19)
RF Location	−0.26° (14.64°)	−0.06° (16.31°)	−1.76° (40.72°)	1.02° (34.45°)	−0.22° (35.71°)	0.10° (35.24°)
RF Size	36.06° (5.05°)	35.99° (5.10°)	29.10° (7.78°)	28.61° (7.29°)	24.24° (4.07°)	24.19° (4.07°)

$N(300,\sigma)$ were σ was varied and negative samples were discarded and re-sampled. A set of experiments were conducted where σ was varied from 0 ms to 2000 ms in increments of 100 ms, resulting in a total of 21 experiments.

Preliminary experiments revealed that longer fixations cause synapses from the input neurons that are active for the current fixation to continue to be strengthened for a longer period of time at the expense of synapses that were potentiated during previous fixations, which are depressed by the continuous global re-normalisation of each output neuron's synaptic weight vector. This process depresses the previously potentiated synapses in favour of the present input to such a degree that output neurons are unable to develop and sustain strong synapses to multiple input patterns with different combinations of eye position and retinal target location corresponding to a particular head-centered target location. A slight alteration of the learning dynamics was therefore introduced to overcome this effect, in the form of an upper bound on individual synapses which prevented long periods of fixation from causing indefinitely long potentiations at the single synapse level at the expense of the rest of the weight vector. Specifically, the learning rule used was

$$\frac{\mathrm{d}w_{ij}}{\mathrm{d}t} = Q(w* - w_{ij})q_i v_j^I \tag{9}$$

where $w*$ is an effective upper bound on each synapse. That is, as approaches $w*$ from below, the rate of change of w_{ij} tends to zero due to the factor $(\mathrm{w}* - w_{ij})$ and hence further potentiation ceases. The weight vector normalization performed at each time step in previous experiments was also included. The parameters for this simulation were identical to those used above, except $w* = 0.15$ and $Q = 2$.

The impact of the variability of fixation duration on the performance of the model was inspected by plotting key summary statistics as a function of fixation duration standard deviation σ in Figure 7. The head-centeredness rate and the average head-centeredness initially decreased with increasing σ. However, the head-centeredness rate and average head-centeredness eventually stabilised just below 60% and above 0.6, respectively. In particular, the head-centeredness rate of the trained model was well above the untrained rate of ~22%. Among head-centered neurons the average head-centeredness and receptive field size both remained very stable across the full range of standard deviations σ, and also well above and below corresponding values in the untrained condition respectively. Finally there was coverage, at no less than ~0.95, across the full range of explored standard deviations. In summary, these results showed that model performance was robust to variability in fixation duration across a wide range of σ.

Discussion

We have hypothesised that head-centered visual neurons might develop in the primate parietal cortex through visually-guided learning by combining trace learning (Foldiak 1991) with the natural statistics of eye and head movement (Freedman and Sparks 1997; Einhäuser et al. 2007). Specifically, if a primate adjusts its gaze by moving its eyes more frequently than its head, then the visual input signals corresponding to a visual target situated at a particular head-centered location will tend to be clustered together in time. In this case, a trace learning rule will

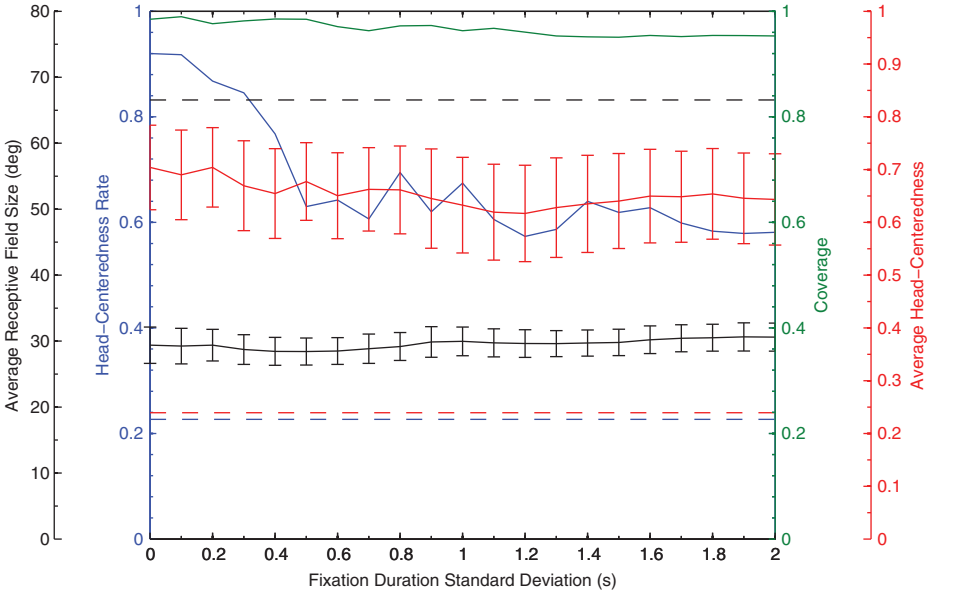

Figure 7. Population summary statistics of the response properties of output neurons from experiments in which the fixation durations during training are continually sampled from a normal distribution $N(300\,\text{ms}, \sigma)$. The dashed lines show the corresponding quantity in the untrained model.

encourage individual postsynaptic neurons to bind these visual input patterns together and thereby develop head-centered visual responses.

The aim of this paper was to extend the idealized experiments in (Mender and Stringer 2013) with more biologically and ecologically constrained model features, thus lending further support to the viability of the model. A range of more realistic model features were successfully accommodated without further difficulty. These model extensions included increasing the number of head-centered target training locations, implementing more realistic stimulus dynamics during training, training on multiple simultaneously visible targets and introducing fixation duration variability. This work is the first time a model of the development of head-centered visual neurons has been examined along these dimensions.

It was found that the model could be trained with visual targets presented along the full continuum of head-centered space, and indeed the performance of the model improved as the density of training locations increased. Previous supervised models have only been trained on a relatively small number of training locations. This showed that, despite the potential for destructive interaction effects between the overlapping representation of adjacent head-centered locations, the self-organization was robust and improved as training conditions became more ecological.

It was found that the model could be trained on visual scenes that always included multiple visual targets, and the output neurons still developed head-centered receptive fields around single training locations. This was a critical finding given that natural visual scenes always contain multiple simultaneously visible targets, and

hence the self-organization of the system must function under these more ecological conditions. For simplicity, only two targets where shown simultaneously, but (Stringer et al. 2007) have studied the problem in further detail for three or more visible stimuli.

The basic model of (Mender and Stringer 2013) employed global re-normalisation of the synaptic weight vectors, but did not incorporate a bound on the magnitude of the individual synaptic weights. It was important to test that this simplest form of network architecture was able to cope well on most of the ecological tests in this current study. However, it was found that introducing variability into the fixation durations could undermine the successful self-organization of head-centered output neurons in this model. This was because neurons then forgot past input patterns in favour of the most recent one presented for a relatively long period of time. To remedy this, one approach that we explored was to introduce an upper bound on individual synapses, which prevented long periods of fixation from causing indefinitely long potentiations at the single synapse level at the expense of the rest of the weight vector. This allowed the model to successfully self-organise with relatively long fixation durations. Moreover, the new model with bounds on individual synaptic weights should perform well on the other ecological tests for the same reasons as the basic model. However, we do not discount that there may be alternative solutions to the problem of variable fixation durations.

It was found that departing from the ideal spatio-temporal training dynamics, in which the eyes move while the head remains stationary, still produced a significant proportion of head-centered output neurons. This finding was particularly critical given that natural eye and head-movements, as well as the dynamics of visual objects in the world, do not conform to the strict training regime previously explored. The result shows the robustness of the underlying principle of temporal association by trace learning, and that it may still function even though the majority of the visual training is dominated by stimulus dynamics radically different from what is actually required for the system to self-organize properly. However, there must of course continue to be some portion of the visual training that follows the stimulus dynamics needed for trace learning to operate, that is, with the eyes moving while the head remains stationary. But this is an entirely realistic requirement of the model.

Nevertheless, the simplified stimulus dynamics used in these simulations were still substantially less rich than what would be observed under natural conditions. Both the retinal and eye position spaces in these simulations were one dimensional, and including a second dimension in both could introduce more complex dynamics. Perhaps more importantly, there was no relationship between the movement dynamics of the head, eyes and the visual targets that the model was presented with. However, this is certainly not the case under natural conditions considering that animals frequently reorient their gaze to acquire visual targets. In future work, it would be valuable to expose a two dimensional version of the model to the eye and head movements recorded from a human subject while viewing a natural visual environment.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the article.

References

Andersen R, Bracewell R, Barash S, Gnadt J, Fogassi L. 1990. Eye position effects on visual, memory, and saccade-related activity in areas LIP and 7a of macaque. The Journal of Neuroscience 10:1176–1196.

Dayan P, Abbott LF. 2001. Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems. Cambridge, MA: MIT Press. ISBN 0-262-04199-5.

Einhäuser W, Schumann F, Bardins S, Bartl K, Böning G, Schneider E, König P. 2007. Human eye-head coordination in natural exploration. Network: Computation in Neural Systems 18:267–297.

Foldiak P. 1991. Learning invariance from transformation sequences. Neural Computation 3:194–200.

Freedman EG, Sparks DL. 1997. Eye-head coordination during head-unrestrained gaze shifts in rhesus monkeys. Journal of Neurophysiology 77:2328–2348.

Galletti C, Battaglini P, Fattori P. 1995. Eye position influence on the parieto-occipital area po of the macaque monkey. European Journal of Neuroscience 7:2486–2501.

Mazzoni P, Andersen RA, Jordan MI. 1991. A more biologically plausible learning rule for neural networks. Proceedings of the National Academy of Sciences 88:4433–4437.

Mender BM, Stringer SM. 2013. A model of self-organizing head-centered visual responses in primate parietal areas. PloS one 8:e81406.

Pouget A, Sejnowski TJ. 1997. Spatial transformations in the parietal cortex using basis functions. Journal of Cognitive Neuroscience 9:222–237.

Rolls ET, Deco G. 2002. Computational neuroscience of vision. Oxford: Oxford University Press.

Rolls ET, Treves A. 1998. Neural Networks and Brain Function. Oxford: Oxford University Press.

Spratling MW. 2009. Learning posture invariant spatial representations through temporal correlations. Autonomous Mental Development, IEEE Transactions on 1:253–263.

Stringer SM, Rolls ET. 2000. Position invariant recognition in the visual system with cluttered environments. Neural Networks 13:305–315.

Stringer SM, Rolls ET. 2008. Learning transform invariant object recognition in the visual system with multiple stimuli present during training. Neural Networks 21:888–903.

Stringer SM, Perry G, Rolls ET, Proske J. 2006. Learning invariant object recognition in the visual system with continuous transformations. Biological Cybernetics 94:128–142.

Stringer S, Rolls E, Tromans J. 2007. Invariant object recognition with trace learning and multiple stimuli present during training. Network: Computation in Neural Systems 18:161–187.

Xing J, Andersen RA. 2000. Models of the posterior parietal cortex which perform multimodal integration and represent space in several coordinate frames. Journal of Cognitive Neuroscience 12:601–614.

Zipser D, Andersen RA. 1988. A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. Nature 331:679–684.