BRAIN RESEARCH ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Available online at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/brainres

Research Report

Natural image sequences constrain dynamic receptive fields and imply a sparse code ☆

Chris Häusler[a,c,*,1], Alex Susemihl[b,c,1], Martin P. Nawrot[a,c]

[a]Neuroinformatics and Theoretical Neuroscience Group, Freie Universität Berlin, Germany
[b]Methods of Artificial Intelligence Group, Berlin Institute of Technology, Germany
[c]Bernstein Center for Computational Neuroscience Berlin, Germany

ARTICLE INFO

Article history:
Accepted 31 July 2013

Keywords:
Autoencoding
Lifetime sparseness
Machine learning
Population sparseness
Restricted Boltzmann Machine
Visual cortex

ABSTRACT

In their natural environment, animals experience a complex and dynamic visual scenery. Under such natural stimulus conditions, neurons in the visual cortex employ a spatially and temporally sparse code. For the input scenario of natural still images, previous work demonstrated that unsupervised feature learning combined with the constraint of sparse coding can predict physiologically measured receptive fields of simple cells in the primary visual cortex. This convincingly indicated that the mammalian visual system is adapted to the natural spatial input statistics. Here, we extend this approach to the time domain in order to predict dynamic receptive fields that can account for both spatial and temporal sparse activation in biological neurons. We rely on temporal restricted Boltzmann machines and suggest a novel temporal autoencoding training procedure. When tested on a dynamic multi-variate benchmark dataset this method outperformed existing models of this class. Learning features on a large dataset of natural movies allowed us to model spatio-temporal receptive fields for single neurons. They resemble temporally smooth transformations of previously obtained static receptive fields and are thus consistent with existing theories. A neuronal spike response model demonstrates how the dynamic receptive field facilitates temporal and population sparseness. We discuss the potential mechanisms and benefits of a spatially and temporally sparse representation of natural visual input.

1. Introduction

Physiological and theoretical studies have argued that the sensory nervous systems of animals are evolutionarily adapted to their natural stimulus environment (for review see Reinagel, 2001). The question of how rich and dynamic natural stimulus conditions determine single neuron response properties and the functional network connectivity in

*Corresponding author at: Neuroinformatics and Theoretical Neuroscience Group, Freie Universität Berlin, Germany. Fax: +49 3083857298.
E-mail addresses: chris.hausler@bccn-berlin.de (C. Häusler), alex.susemihl@bccn-berlin.de (A. Susemihl), martin.nawrot@fu-berlin.de (M.P. Nawrot).

[1]These authors contributed equally to the published work.

0006-8993/$ - see front matter © 2013 The Authors. Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.brainres.2013.07.056

mammalian sensory pathways has thus become an important focus of interest for theories of sensory coding (for review see Simoncelli and Olshausen, 2001; Olshausen et al., 2004).

For a variety of animal species and for different modalities it has been demonstrated that single neurons respond in a temporally sparse manner (Reinagel, 2001; Jadhav et al., 2009; Olshausen et al., 2004; Hromádka et al., 2008) when stimulated with natural time-varying input. In the mammal this is intensely studied in the visual (Dan et al., 1996; Vinje and Gallant, 2000; Reinagel and Reid, 2002; Yen et al., 2007; Maldonado et al., 2008; Haider et al., 2010; Martin and Schröder, 2013) and the auditory (Hromádka et al., 2008; Chen et al., 2012; Carlson et al., 2012) pathway as well as in the rodent whisker system (Jadhav et al., 2009; Wolfe et al., 2010). Sparseness increases across sensory processing levels and is particularly high in the neocortex. Individual neurons emit only a few spikes positioned at specific instances during the presentation of a time-varying input. Repeated identical stimulations yield a high reliability and temporal precision of responses (Herikstad et al., 2011; Haider et al., 2010). Thus, single neurons focus only on a highly specific spatio-temporal feature from a complex input scenario.

Theoretical studies addressing the efficient coding of natural images in the mammalian visual system have been very successful. In a ground breaking study, Olshausen et al. (1996) learned a dictionary of features for reconstructing a large set of natural still images under the constraint of a sparse code to obtain receptive fields (RFs), which closely resembled the physiologically measured RFs of simple cells in the mammalian visual cortex. This approach was later extended to the temporal domain by van Hateren and Ruderman (1998), learning rich spatio-temporal receptive fields directly from movie patches. In recent years, it has been shown that a number of unsupervised learning algorithms, including the denoising Autoencoder (dAE) (Vincent et al., 2010) and the Restricted Boltzmann Machine (RBM) (Hinton and Salakhutdinov, 2006; Hinton et al., 2012; Mohamed et al., 2011), are able to learn structure from natural stimuli and that the types of structure learnt can again be related to cortical RFs as measured in the mammalian brain (Saxe et al., 2011; Lee et al., 2008, 2009).

Considering that sensory experience is per se dynamic and under the constraint of a temporally sparse stimulus representation at the level of single neurons, how could the static RF model, i.e. the learned spatial feature, extend into the time domain? Here we address this question with an unsupervised learning approach using RBMs as a model class. Building on an existing model, the Temporal Restricted Boltzmann Machine (TRBM) introduced by Sutskever and Hinton (2007), we introduce a novel learning algorithm with a temporal autoencoding approach to train RBMs on natural multi-dimensional input sequences. For validation of the method, we test the performance of our training approach on a reference dataset of kinematic variables of human walking motion and compare it against the existing TRBM model and the Conditional RBM (CRBM) as a benchmark (Taylor et al., 2007). As an application of our model, we train the TRBM using temporal autoencoding on natural movie sequences and find that the neural elements develop dynamic RFs that express smooth transitions, i.e. translations and rotations, of the static receptive field model. Our model neurons account for spatially and temporally sparse activities during stimulation with natural image sequences and we demonstrate this by simulation of neuronal spike train responses driven by the dynamic model responses. Our results propose how neural dynamic RFs may emerge naturally from smooth image sequences.

2. Results

We outline a novel method to learn temporal and spatial structure from dynamic stimuli – in our case smooth image sequences – with artificial neural networks. The hidden units (neurons) of these generative models develop dynamic RFs that represent smooth temporal evolutions of static RF models that have been described previously for natural still images. When stimulated with natural movie sequences the model units are activated sparsely, both in space and time. A point process model translates the model's unit activation into sparse neuronal spiking activity with few neurons being active at any given point in time and sparse single neuron firing patterns.

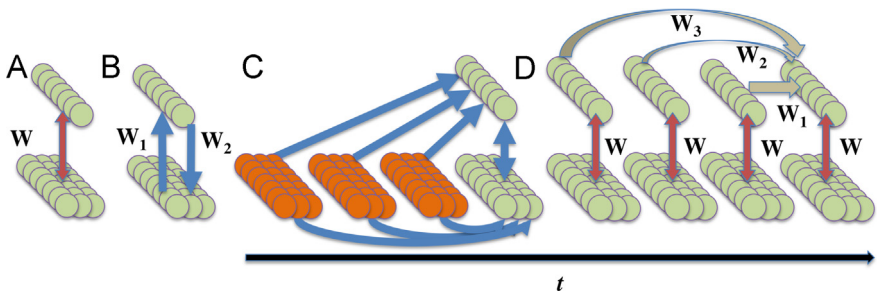

Fig. 1 – Described model architectures: (A) Autoencoder; (B) RBM; (C) Conditional RBM and (D) Temporal RBM. In the CRBM (subfigure C; see also Section 4), there is a *hidden* layer only at the current sample time whose activation is defined by weights connecting the current as well as previous activations of the *visible* layer. The TRBM (subfigure D) has a *hidden* layer instantiation for each sample time within the models delay dependency and the temporal evolution of the model is defined by lateral connections between the *hidden* units of consecutive time steps.

Please cite this article as: Häusler, C., et al., Natural image sequences constrain dynamic receptive fields and imply a sparse code. Brain Research (2013), http://dx.doi.org/10.1016/j.brainres.2013.07.056

2.1. The model

We rely on the general model class of RBMs (see Section 4.1). The classic RBM is a two layer artificial neural network with a *visible* and a *hidden* layer used to learn representations of a dataset in an unsupervised fashion (Fig. 1A). The units (neurons) in the *visible* and those in the *hidden* layers are all-to-all connected via symmetric weights and there is no connectivity between neurons within the same layer. The input data, in our case natural images, activate the units of the *visible* layer. This activity is then propagated to the *hidden* layer where each neuron's activity is determined by the input data and by the weights **W** connecting the two layers. The weights define each hidden neuron's filter properties or its RF, determining its preferred input.

Whilst the RBM has been successfully used to model static data, it lacks in the ability to explicitly represent the temporal evolution of a continuous dataset. The CRBM (Fig. 1C) and TRBM (Fig. 1D) are both temporal extensions of the RBM model, allowing the hidden unit activations to be dependent on multiple samples of a sequential dataset. The models have a *delay* parameter which is used to determine how long the integration period on a continuous dataset is.

The CRBM has an instantiation of the *visible* layer for each sample time within the model's delay range, each of which is connected directly to the single *hidden* layer at the current sample point. In the TRBM (Fig. 1D; see also Fig. 4.1) the temporal dependence is modelled by a set of weights connecting the *hidden* layer activations at previous steps in the sequence to the current hidden layer representation. The TRBM and CRBM have proven to be useful in the modelling of temporal data, but each again has its drawbacks. The CRBM does not separate the representations of *form* and *motion*. Here we refer to *form* as the RF of a hidden unit in one sample of the dataset and *motion* as the evolution of this feature over multiple sequential samples. This drawback makes it difficult to interpret the features learnt by the CRBM over time as the two modalities are mixed. The TRBM explicitly separates representations of *form* and *motion* by having dedicated weights for the *visible* to *hidden* layer connections (*form*) and for the temporal evolution of these features (*motion*). Despite these benefits, the TRBM has proven quite difficult to train due to the intractability of its probability distribution (see Fig. 4).

In this work we develop a new approach to training Temporal Restricted Boltzmann Machines that we call Temporal Autoencoding (we refer to the resulting TRBM as an autoencoded TRBM or aTRBM) and investigate how it can be applied to modelling natural image sequences. The aTRBM adds an additional step to the standard TRBM training, leveraging a denoising Autoencoder to help constrain the temporal weights in the model. Table 1 provides an outline of the training procedure whilst more details can be found in Section 4.1.3.

In the following sections we compare the filters learnt by the aTRBM and CRBM models on natural image sequences and show that the aTRBM is able to learn spatially and temporally sparse filters having response properties in line with those found in neurophysiological experiments.

2.2. Learning temporal filters from natural image sequences

We have trained a CRBM and an aTRBM on natural image sequence data taken from the Hollywood2 dataset introduced in Marszalek et al. (2009), consisting of a large number of snippets from various Hollywood films. From the dataset, 20×20 pixel patches are extracted in sequences 30 frames long. Each patch is contrast normalized (by subtracting the mean and dividing by the standard deviation) and ZCA whitened (Bell and Sejnowski, 1997) to provide a training set of approximately 350,000 samples. The aTRBM and CRBM models, each with 400 hidden units and a temporal dependency of 3 frames, are trained initially for 100 epochs on static frames of the data to initialize the static weights **W** and then until convergence on the full temporal sequences. Full details of the models' architecture and training approaches are given in the Experimental procedures section.

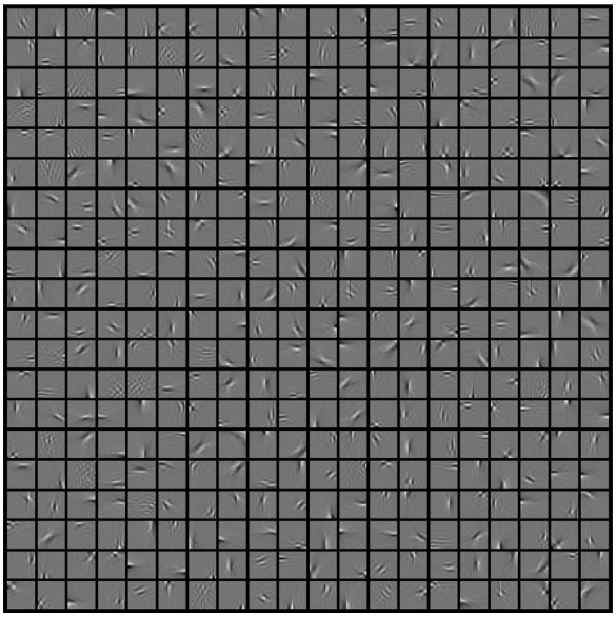

Fig. 2 – Static filters learned by the aTRBM on 20×20 image patches. Note the mostly Gabor like filters of varying orientation and frequency selectivity.

Table 1 – Autoencoded TRBM training steps.

Step	Action
1. Static RBM training	Constrain the static weights $\mathbf{w}$ using CD on single frame samples of the training data
2. Temporal autoencoding	Constrain the temporal weights $\mathbf{w}_1$ to $\mathbf{w}_d$ using a denoising autoencoder on multi-frame samples of the data
3. Model finalization	Train all model weights together using CD on multi-frame samples of the data

Please cite this article as: Häusler, C., et al., Natural image sequences constrain dynamic receptive fields and imply a sparse code. Brain Research (2013), http://dx.doi.org/10.1016/j.brainres.2013.07.056

2.2.1. *Static RFs*

The static filters learned by the aTRBM through the initial contrastive divergence training can be seen in Fig. 2 (note that the static filters are pre-trained in the same way for the CRBM and aTRBM, therefore the filters are equivalent). We obtain Gabor-like patch filters resembling simple cell RFs in V1, reproducing the typical result for a variety of methods (see Introduction), statistics of which can be seen in Fig. 3. The RFs of the hidden units are spatially located across the entire image patch with some distinct clustering along the borders (Fig. 3A). In 2D Fourier space (Fig. 3B) one can see a good coverage of the space, representing frequency and direction selectivity, both these results being in agreeance with those found in similar studies (see Cadieu and Olshausen, 2012; Bell and Sejnowski, 1997, for example). The filters also display a preference for cardinal (horizontal and vertical) orientations (Fig. 3C), a phenomenon that has often been reported in electrophysiological experiments of primary visual cortex (e.g. Wang et al., 2003; Coppola et al., 1998).

2.2.2. *Dynamic RFs*

We then analysed how the static filters are connected through the temporal weights learned during autoencoder training by visualizing their evolution over time. The filters

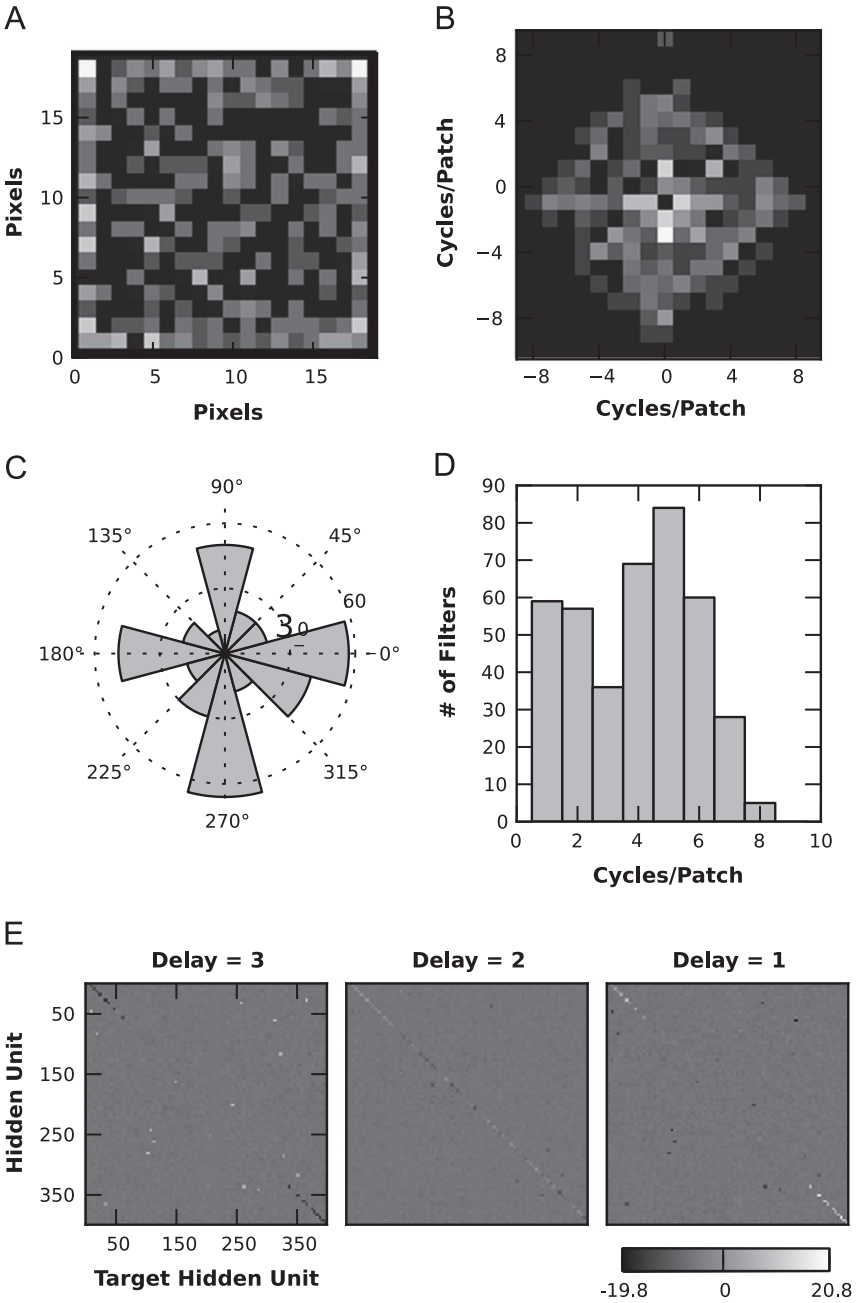

Fig. 3 – Static filter statistics – aTRBM: (A) histogram of the filters spatial location; (B) histogram of the filters spatial frequency; (C) histogram of the filters preferred direction (showing a clear preference for cardinal directions) and (D) frequency. (E) Visualization of the temporal transition weights for 3 time delays for the aTRBM. Note the strong self-excitation at *delay*=1 and self-inhibition at *delay*=3.

Please cite this article as: Häusler, C., et al., Natural image sequences constrain dynamic receptive fields and imply a sparse code. Brain Research (2013), http://dx.doi.org/10.1016/j.brainres.2013.07.056

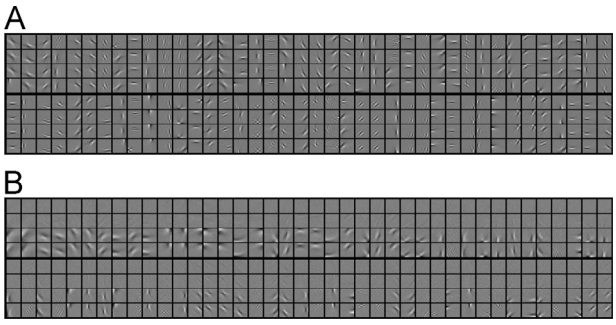

Fig. 4 – Dynamic RFs. 80 out of 400 hidden units with the highest temporal variation from an aTRBM (A) and a CRBM (B). For the pre-trained TRBM, we plot the most active units as described in the text. Each group of 4 images represents the temporal filter of one hidden unit with the lowest patch representing time *t* and the 3 patches above representing each of the delay steps in the model. The units are displayed in two rows of 40 columns with 4 filters, with the temporal axis going from top to bottom.

discussed were learned by the aTRBM (see Eq. (1)) with our training algorithm described in Section 4.1.3. To visualize the dynamic RF of a hidden unit we clamped the activation of that unit to 1 and set all other units to be inactive in the most delayed layer of the aTRBM. We then proceeded to sample from the distribution of all other hidden layers and chose the most active units in every delay. This is shown in Fig. 4. We have shown the most active units when a hidden unit is active for the 80 units with highest temporal variation among the subsequent filters. This, however, only gives us a superficial look into the dynamics of the RFs. One way to look further is to consider the n most active units at the second-furthest delay and then sequentially clamp each of these to an active state and look at the resulting activations in the remaining layers. If one does this sequentially, we are left with a tree of active units, 1 at time $t-T$, n at time $t-(T-1)$, and n^T at time t. We can then look at what these units code for. We have performed this procedure with two hidden units, and to visualize what they code for we have plotted the center of mass of the filters in frequency and position space. This is shown in Fig. 5.

Visualizing the temporal RFs learnt by the CRBM is simpler than for the aTRBM. We display the weight matrix $\mathbf{W}$ and the temporal weights $\mathbf{W}_1$ to $\mathbf{W}_d$ for each hidden unit directly as a projection into the visible layer (a 20×20 patch). This shows the temporal dependence of each hidden unit on the past visible layer activations and is plotted with time running from top to bottom in Fig. 4B. The aTRBM learns richer filter dynamics with a longer temporal dependency, whereas the CRBM only seems to care about the visible layers at times t and $t-1$, possibly because most of the variation is captured by the visible-to-visible weights. The temporal profile of excitation versus inhibition for the aTRBM can also be seen from the profile of the connectivity matrix between its hidden units. This is shown in Fig. 3E and one can note a transition from self-excitation at $delay=1$ to self-inhibition at $delay=3$.

In Fig. 5 we analyse the filter histories of the aTRBM for $n=3$ and visualize for two of the *hidden* layer units, their preference in image space, frequency and direction.

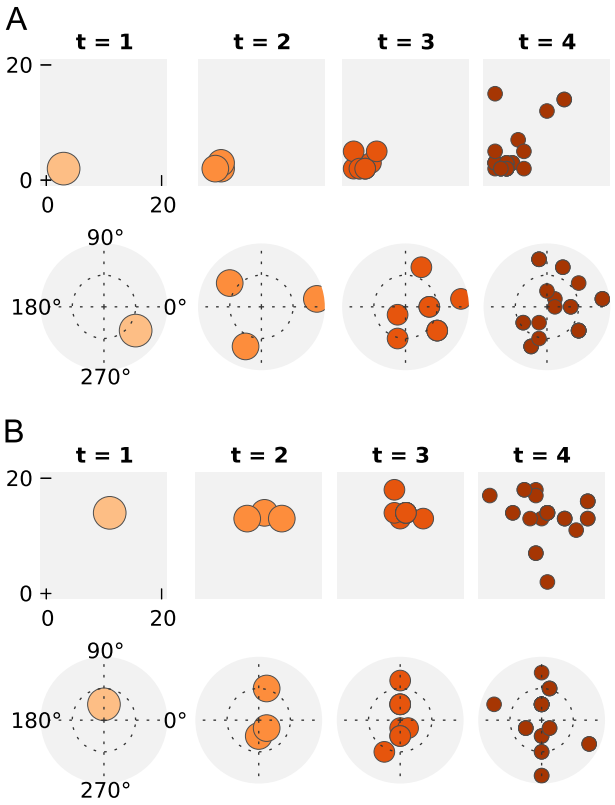

Fig. 5 – Spatial and angular evolution of two hidden units in the aTRBM (subfigures A and B). The upper row shows the center of each units receptive field in pixel space for the most active units in the temporal evolution of one unit. The lower row shows the strongest frequency component of the filters for this same evolution. The unit in subfigure (A) shows a clear spatial preference but is orientation agnostic whilst the unit in subfigure (B) is less spatially selective but shows a clear preference for vertically oriented filters.

For the unit in Fig. 5A there is a clear selectivity for spatial location over its temporal evolution and activations remain spatially localized. In contrast there is no apparent preference for orientation. The unit depicted in Fig. 5B, on the other hand, displays strong orientation selectivity, but the spatial selectivity is not accentuated. These results are representative of the population and provide evidence for preferential connectivity between cells with similar RFs, a finding that is supported by a number of experimental results in V1 (Bosking et al., 1997; Field and Hayes, 2004).

2.3. *The dynamic RF model facilitates sparse coding*

The temporal evolution of the spatial filter structure expressed by single units in the dynamic RF model (Figs. 4 and 5) renders individual units to be selective to a specific spatio-temporal structure of the input within their classical RF. This increased stimulus specificity in comparison to a static RF model implies an increased sparseness of the units' activation. To test this hypothesis we quantified temporal and spatial sparseness for both model approaches.

Please cite this article as: Häusler, C., et al., Natural image sequences constrain dynamic receptive fields and imply a sparse code. Brain Research (2013), http://dx.doi.org/10.1016/j.brainres.2013.07.056

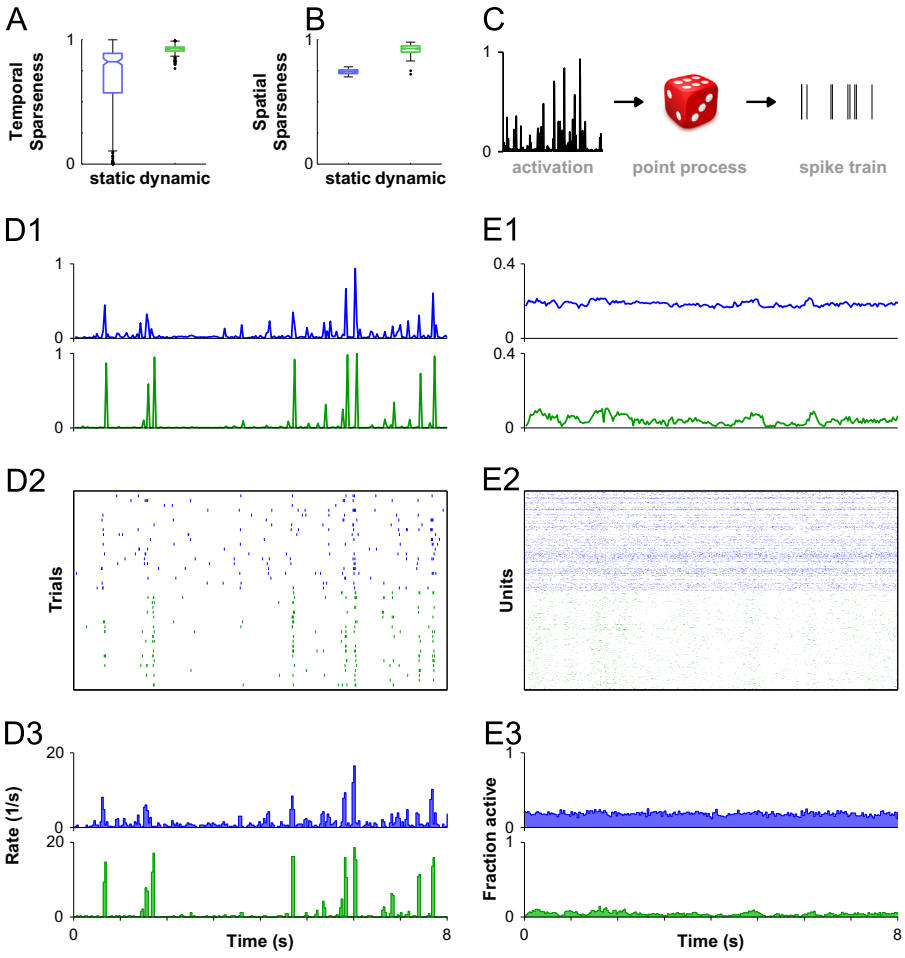

Fig. 6 – Temporal and spatial sparseness of neuronal activity for static and dynamic RF responses. (A) Temporal sparseness measured in 400 hidden layer units during 30 s of video stimulation is significantly larger for the dynamic (right) than the static (left) RF model ($P<10^{-5}$; Wilcoxon signed rank test). (B) Spatial sparseness measured across all 400 neurons is significantly increased ($P<10^{-5}$; Wilcoxon signed rank test) in the dynamic (right) RF model as compared to the static RF model (left). (C) Sketch of cascade model for spike train generation. During video stimulation the activation curve of a hidden layer neuron (left) expresses the deterministic probability of being active in each frame. A stochastic point process model (center) generates action potentials (right) according to a time-varying intensity proportional to the activation curve. (D1–D3) Temporal sparseness during 8 s of video stimulation. (D1) Activation curve of one hidden neuron for the static RF (blue) and the dynamic RF (green) model with a temporal sparseness of $S=0.82$ and $S=0.94$, respectively. (D2) Repeated point process realizations ($n=20$) using the activation curves in (D1). (D3) Firing rate estimated as time histogram from 100 repetitions for static (blue) and dynamic (green) RF model. (E1–E3) Spatial sparseness in the population of hidden layer neurons during video stimulation. (E1) Average activation curves of hidden layer units for the static (blue) and dynamic (green) RF model. (E2) Spike trains of $N=50$ hidden layer neurons when using the static (red) or dynamic (blue) RF model. (E3) The fraction of active neurons per video frame in the total population of 400 hidden units is considerably smaller for the dynamic RF model.

2.3.1. *Temporal sparseness*

We measured temporal sparseness of the single unit activation h using the well established sparseness index S (equation (2)) introduced by Willmore and Tolhurst (2001) and described in Section 4.2.1. The higher the value of S for one particular unit, the more peaked is the temporal activation profile $h(t)$ of this unit. The lower the value of S, the more evenly distributed are the activation values $h(t)$. The quantitative results across the population of 400 hidden units in our aTRBM model are summarized in Fig. 6A. As expected, units are temporally sparser when the dynamic RF is applied with a mean sparseness index of 0.92 (median: 0.93) compared to the mean of 0.69 (median: 0.82) for the static RF. This is also reflected in the activation curves for one example unit shown in Fig. 6D1 for the static RF (blue) and the dynamic RF (green) recorded during the first 8 s of video input.

In the nervous system temporally sparse stimulus encoding finds expression in stimulus selective and temporally structured single neuron firing patterns where few spikes are emitted at specific instances in time during the presentation of a time varying stimulus (see Section 1). In repeated stimulus presentations the temporal pattern of action potentials is typically repeated with high reliability (e.g. Herikstad et al., 2011). In order to translate the continuous activation variable of the hidden units in our aTRBM model into spiking activity we used the cascade model depicted in Fig. 6C and

Please cite this article as: Häusler, C., et al., Natural image sequences constrain dynamic receptive fields and imply a sparse code. Brain Research (2013), http://dx.doi.org/10.1016/j.brainres.2013.07.056

described in Section 4.2.2. The time-varying activation curve (Fig. 6D1) is used as deterministic intensity function of a stochastic point process model. This allows us to generate repeated stochastic point process realizations, i.e. single trial spike trains, as shown for the example unit in Fig. 6D2. Clearly, the repeated simulation trials based on the dynamic RF activation (green) exhibit a spiking pattern, which is temporally sparser than the spiking pattern that stems from the static RF activation (blue). This also finds expression in the time histogram of the trial-averaged firing rate shown in Fig. 6D3. The firing rate is more peaked in the case of the dynamic RF, resembling the deterministic activation curve in Fig. 6D1.

2.3.2. *Spatial sparseness*

Spatial sparseness (also termed population sparseness) refers to the situation where only a small number of units are significantly activated by a given stimulus. In the natural case of time-varying stimuli this implies a small number of active neurons in any small time window while the rest of the neuron population expresses a low baseline activity. Again, we use S (Eq. (2)) to quantify spatial sparseness from the population activation **h** of hidden neurons and for each time step separately. The results depicted in Fig. 6B show a significantly higher spatial sparseness when the dynamic RF was applied with a mean (median) of 0.92 (0.93) as compared to the static RF with a mean (median) of 0.74 (0.74).

We demonstrate how the spatial sparseness for the static and the dynamic RF model in the population of hidden units affects spiking activity using our cascade point process model. Fig. 6E2 shows the simulated spiking activity of all 400 neurons based on the activation $\mathbf{h}(t)$ of the hidden neurons during 8 s of recording. Overall the static RF (blue) results in higher firing rates. The stimulus representation in the ensemble spike train appears more dense for the static RF (blue) than in the case of a dynamic RF (green). As shown in Fig. 6E3, fewer neurons were active at any given point in time when they were driven by the dynamic RF model.

3. Discussion

We suggested a novel approach to unsupervised learning of spatio-temporal structure in multi-dimensional time-varying data. We first define the general topology of an artificial neural network (ANN) as our model class. Through a number of structural constraints and a machine learning approach to train the model parameters from the data, we arrive at a specific ANN which is biologically relevant and is able to produce activations for any given temporal input (Section 2.1). We then extend this ANN with a Computational Neuroscience based cascade model and use this to generate trial variable spike trains (Section 2.3).

The proposed aTRBM model integrates the recent input history over a small number of discrete time steps. This model showed superior performance to other models on a recognized benchmark dataset. When trained on natural videos that represent smooth sequences of natural images the units in the hidden layer developed dynamic receptive fields that retain the properties of the static receptive field and represent smooth temporal transitions of their static receptive field structure. This time-extension of the previously obtained static receptive fields increase the input

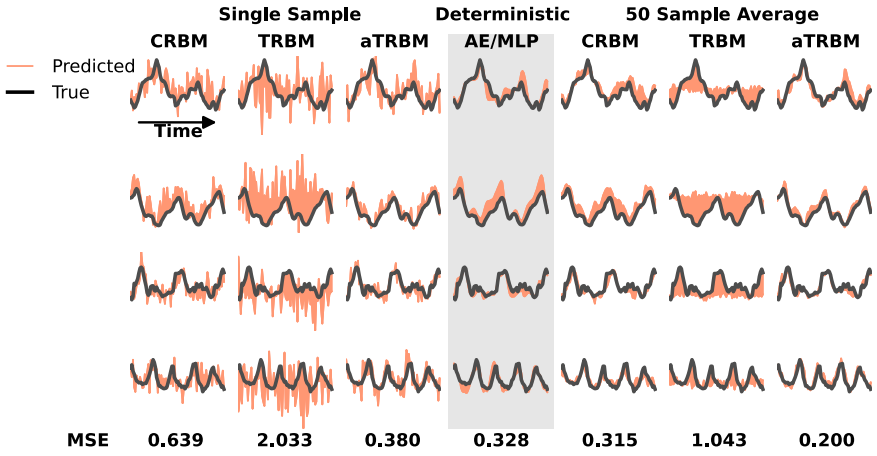

Fig. 7 – CRBM, TRBM, aTRBM and AE/MLP are used to fill in data points from motion capture data (Taylor et al., 2007). Four random dimensions of the motion data are shown along with their model reconstructions from a single trial (three leftmost columns), deterministically (middle column, grey), and as an average of 50 generated trials (three rightmost columns). At the bottom of each column, one can see the Mean Squared Error (MSE) of the reconstruction over all 49 dimensions of the entire 1000 sample test data. The aTRBM is the best performer of the single trial predictors, producing a lower MSE than the CRBM and TRBM. The deterministic AE/MLP has marginally better MSE performance than the aTRBM, although at the cost of no longer being a generative model. We find, however, that if one generates 50 single trial predictions from the aTRBM model and then takes the average of these, the MSE is reduced ever further, allowing the aTRBM to far outperform the AE/MLP. From this point of view, the aTRBM is the more advantageous model in the respect that it can generate non-deterministic single trial predictions, and if one is interested in reducing the MSE as far as possible, can be averaged over a number of trials, thereby reducing the single trial variation and increasing the predictor performance.

Please cite this article as: Häusler, C., et al., Natural image sequences constrain dynamic receptive fields and imply a sparse code. Brain Research (2013), http://dx.doi.org/10.1016/j.brainres.2013.07.056

selectivity of each hidden unit. Consequently, each hidden unit is activated in a highly sparse manner by only specific spatio-temporal input scenarios.

3.1. *Temporal autoencoding model*

We have introduced a new training method for TRBMs called Temporal Autoencoding and validated it by showing a significant performance increase in modelling and generation from a sequential human motion capture dataset (Fig. 7). The gain in performance from the standard TRBM to the pre-trained aTRBM model, which are both structurally identical, suggests that our approach of autoencoding the temporal dependencies gives the model a more meaningful temporal representation than is achievable through contrastive divergence training alone. We believe the inclusion of autoencoder training in temporal learning tasks will be beneficial in a number of problems, as it enforces the causal structure of the data on the learned model.

We have shown that the aTRBM is able to learn high level structure from natural movies and account for the transformation of these features over time. The statistics of the static filters resemble those learned by other algorithms, namely Gabor like patches showing preferential orientation of the filters along cardinal directions (Fig. 2). The distribution of preferred position, orientation and frequency (Fig. 3) is in accordance with results previously found by other methods (e.g. Cadieu and Olshausen, 2008; Bell and Sejnowski, 1997), and the simple cell like receptive fields and cardinal selectivity is supported by neurophysiological findings in primary visual cortex (Wang et al., 2003; Coppola et al., 1998). Importantly the temporal connectivity expressed in the weights $\mathbf{W}_M$ learned by the model is also qualitatively similar to the pattern of lateral connections in this brain area. Preferential connection between orientation-selective cells in V1 with similar orientation has been reported in higher mammals (Bosking et al., 1997; Field and Hayes, 2004; Van Hooser, 2007). These lateral connections are usually thought to underlie contour integration in the visual system. Here they arise directly from training the aTRBM model to reproduce the natural dynamics of smoothly changing image sequences. One could say that, in an unsupervised fashion, the model learns to integrate contours directly from the dataset.

The aTRBM presented here can be easily embedded into a deep architecture, using the same training procedure in a greedy layer-wise fashion. This might allow us to study the dynamics of higher-order features (i.e. higher order receptive fields) in the same fashion as was done here for simple visual features. In this way one could envisage applications of our approach to pattern recognition and temporal tasks, such as object tracking or image stabilization.

3.2. *The dynamic RF is a potential mechanism of sparse stimulus encoding*

There is strong evidence that encoding of natural stimuli in sensory cortices – specifically in the visual and auditory system – is sparse in space and time (see Section 1). Sparse coding seems to be a universal principle widely employed both in vertebrate and invertebrate nervous systems and it is thought to reflect the sparsity of natural stimulus input (Vinje and Gallant, 2000; Olshausen et al., 2004; Zetzsche and Nuding, 2005). Deciphering the neuronal mechanisms that underlie sparse coding at the level of cortical neurons is a topic of ongoing research.

Population sparseness critically depends on the network topology. An initially dense code in a smaller population of neurons in the sensory periphery is transformed into a spatially sparse code by diverging connections onto a much larger number of neurons in combinations with highly selective and possibly plastic synaptic contacts. This is particularly well studied in the olfactory system of insects where feed-forward projections from the antennal lobe diverge onto a much larger number of Kenyon cells in the mushroom body with random and weak connectivity (Caron et al., 2013) and thereby translate a dense combinatorial code in the projection neuron population into a sparse code in the Kenyon cell population (Jortner et al., 2007; Huerta and Nowotny, 2009). Also in the mammalian visual system the number of retinal cells at the periphery, which employ a relatively dense code, is small compared to the cortical neuron population in the primary visual cortex (Olshausen et al., 2004). Another important mechanism responsible for spatial sparseness is global and structured lateral inhibition that has been shown to increase population sparseness in the piriform cortex (Poo and Isaacson , 2009) and to underlie non-classical receptive fields in the visual cortex (Haider et al., 2010).

A network architecture of diverging connections and mostly weak synapses is reflected in the RBM models introduced here (see Section 4 and Fig. 1). Initially an all-to-all connection between the units in the input and in the hidden layer is given, but due to the sparsity constraint most synaptic weights become effectively zero during training. By this, hidden layer units sparsely mix input signals in many different combinations to form heterogeneous spatial receptive fields (Fig. 2) as observed in the visual cortex (Reich et al., 2001; Yen et al., 2007; Martin and Schröder, 2013). A novelty of the aTRBM is that the learning of sparse connections between hidden units also applies to the temporal domain resulting in heterogeneous spatio-temporal receptive fields (Fig. 4A). Our spike train simulations (Fig. 6) match the experimental observations in the visual cortex: sparse firing in time and across the neuron population (e.g. Yen et al., 2007; Martin and Schröder, 2013).

Experimental evidence in the visual cortex suggests that temporally sparse responses of single neurons to naturalistic dynamic stimuli show less variability across trials than responses to artificial noise stimuli (Herikstad et al., 2011; Haider et al., 2010). Equally, in the insect olfactory system the temporally sparse stimulus responses in the Kenyon cells have been shown to be highly reliable across stimulus repetitions (Ito et al., 2008). In our model approach, response variability is not affected by the choice of a static or dynamic RF model. The trained aTRBM provides a deterministic activation $\mathbf{h}$ across the hidden units. In the cascade model (Fig. 6C) we generated spike trains according to a stochastic point process model. Thus the trial-to-trial spike count variability in our model is solely determined by the point process stochasticity and is thereby independent of the RF type. Spike frequency adaptation (SFA, Benda and Herz, 2003)

Please cite this article as: Häusler, C., et al., Natural image sequences constrain dynamic receptive fields and imply a sparse code. Brain Research (2013), http://dx.doi.org/10.1016/j.brainres.2013.07.056

is an important cellular mechanism that increases temporal sparseness (Farkhooi et al., 2012; Nawrot, 2012) and at the same time reduces the response variability of single neuron (Chacron et al., 2001; Nawrot et al., 2007; Farkhooi et al., 2009; Nawrot, 2010) and population activity (Chacron et al., 2005; Farkhooi et al., 2011, 2012). Other mechanisms that can facilitate temporal sparseness are feed-forward (Assisi et al., 2007) and feed-back inhibition (Papadopoulou et al., 2011).

3.3. *Why sparse coding?*

Encoding of a large stimulus space can be realized with a dense code or with a sparse code. In a dense coding scheme few neurons encode stimulus features in a combinatorial fashion where each neuron is active for a wide range of stimuli and with varying response rates (stimulus tuning). Dense codes have been described in different systems, prominent examples of which are the peripheral olfactory system of invertebrates and vertebrates (e.g. Friedrich and Laurent, 2004; Wilson et al., 2004; Krofczik et al., 2008; Brill et al., 2013), and the cortical motor control system of primates (e.g. Georgopoulos et al., 1982; Rickert et al., 2009).

In sensory cortices a sparse stimulus representation is evident (see Section 1). Individual neurons have highly selective receptive fields and a large number of neurons is required to span the relevant stimulus space. What are the benefits of a sparse code that affords vast neuronal resources to operate at low spiking rates? We briefly discuss theoretical arguments that outline potential computational advantages of a sparse stimulus encoding.

The first and most comprehensive argument concerns the energy efficiency of information transmission. Balancing the cost of action potential generation relative to the cost for maintaining the resting state with the sub-linear increase of information rate with firing rate in a single neuron leads to an optimal coding scheme where only a small percentage of neurons is active with low firing rates (Levy and Baxter, 1996; Laughlin et al., 2001; Lennie, 2003).

The argument outlined above is limited to the transmission of information and conditioned on the assumption of independent channels. Nervous systems, however, have evolved as information processing systems and information transmission plays only a minor role. Then the more important question is how does sparse coding benefit brain computation? We consider three related arguments. In a spatially sparse code, single elements represent highly specific stimulus features. A complex object can be formed only through the combination of specific features at the next level, a concept that is often referred to as the binding hypothesis (Knoblauch et al., 2001). In this scheme, attentional mechanisms could mediate a perceptual focus of objects with highly specific features by enhancing co-active units and suppressing background activity. In a dense coding scheme, enhanced silencing of individual neurons would have an unspecific effect.

A spatially sparse stimulus representation can facilitate the formation of associative memories (Palm, 1980). A particular object in stimulus space activates a highly selective set of neurons. Using an activity-dependent mechanism of synaptic plasticity allows the formation of stimulus-specific associations in this set of neurons. This concept is theoretically and experimentally well studied in the insect mushroom body where the sparse representation of olfactory stimuli at the level of the Kenyon cells (Perez-Orive et al., 2002; Honegger et al., 2011) is thought to underlie associative memory formation during classical conditioning (Huerta et al., 2004; Huerta and Nowotny, 2009; Cassenaer and Laurent, 2012; Strube-Bloss et al., 2011). This system has been interpreted in analogy to machine learning techniques that employ a strategy of transforming a lower dimensional input space into a higher dimensional feature space to improve stimulus classification (Huerta and Nowotny, 2009; Huerta, 2013; Pfeil et al., 2013).

Theories of temporal coding acknowledge the importance of the individual spike and they receive support from accumulating experimental evidence (e.g. Riehle et al., 1997; Maldonado et al., 2008; Jadhav et al., 2009). Coding schemes that rely on dynamic formation of cell assemblies and exact spike timing work best under conditions of spatially and a temporally sparse stimulus representations and low background activity.

4. Experimental procedures

4.1. *Machine learning methods*

To develop the Temporal Autoencoding training method for Temporal Restricted Boltzmann Machines used in this work, we have extended upon existing work in the field of unsupervised feature learning.

4.1.1. *Existing static models of unsupervised learning*

Two unsupervised learning methods well known within the Machine Learning community, Restricted Boltzmann Machines (RBMs) and Autoencoders (AEs) (Larochelle and Bengio, 2008; Bengio et al., 2007) form the basis of our temporal autoencoding approach. Both are two-layer neural networks, all-to-all connected between the layers but with no intra-layer connectivity. The models consist of a visible and a hidden layer, where the visible layer represents the input to the model whilst the hidden layer's job is to learn a meaningful representation of the data in some other dimensionality. We will represent the visible layer activation variables by v_i, the hidden activations by h_j and the vector variables by $\mathbf{v}=\{v_i\}$ and $\mathbf{h}=\{h_j\}$ where $i=[1..N]$ and $j=[1..S]$ index the individual neurons in the visible and hidden layers, respectively.

Restricted Boltzmann Machines are stochastic models that assume symmetric connectivity between the visible and hidden layers (see Fig. 1A) and seek to model the structure of a given dataset. They are energy-based models, where the energy of a given configuration of activations $\{v_i\}$ and $\{h_j\}$ is given by

$$E_{RBM}(\mathbf{v},\mathbf{h}|\mathbf{W},\mathbf{b}_v,\mathbf{b}_h) = -\mathbf{v}^\top \mathbf{W}\mathbf{h} - \mathbf{b}_v^\top \mathbf{v} - \mathbf{b}_h^\top \mathbf{h},$$

and the probability of a given configuration is given by

$$P(\mathbf{v},\mathbf{h}) = \exp\left(-E_{RBM}(\mathbf{v},\mathbf{h}|\mathbf{W},\mathbf{b}_v,\mathbf{b}_h)\right)/Z(\mathbf{W},\mathbf{b}_v,\mathbf{b}_h),$$

Please cite this article as: Häusler, C., et al., Natural image sequences constrain dynamic receptive fields and imply a sparse code. Brain Research (2013), http://dx.doi.org/10.1016/j.brainres.2013.07.056

where $Z(\mathbf{W},\mathbf{b}_v,\mathbf{b}_h)$ is the partition function. One can extend the RBM to continuous-valued visible variables by modifying the energy function, to obtain the Gaussian-binary RBM

$$E_{RBM}(\mathbf{v},\mathbf{h}|\mathbf{W},\mathbf{b}_v,\mathbf{b}_h) = -\frac{\mathbf{v}^\top}{\sigma^2}\mathbf{W}\mathbf{h} + \frac{\|\mathbf{b}_v-\mathbf{v}\|^2}{2\sigma^2} - \mathbf{b}_h^\top\mathbf{h}.$$

RBMs are usually trained through contrastive divergence, which approximately follows the gradient of the cost function

$$CD_n(\mathbf{W},\mathbf{b}_v,\mathbf{b}_h)) = KL(P_0(\mathbf{v}|\mathbf{W},\mathbf{b}_v,\mathbf{b}_h)||P(\mathbf{v}|\mathbf{W},\mathbf{b}_v,\mathbf{b}_h)) - KL(P_n(\mathbf{v}|\mathbf{W},\mathbf{b}_v,\mathbf{b}_h)||P(\mathbf{v}|\mathbf{W},\mathbf{b}_v,\mathbf{b}_h)),$$

where P_0 is the data distribution and P_n is the distribution of the visible layer after n MCMC steps (Carreira-Perpinan and Hinton, 2005). The function CD_n gives an approximation to maximum-likelihood (ML) estimation of the weight matrix $\mathbf{w}$. Maximizing the marginal probability $P(\{\mathbf{v}\}_D|\mathbf{W},\mathbf{b}_v,\mathbf{b}_h)$ of the data $\{\mathbf{v}\}_D$ in the model leads to a ML-estimate which is hard to compute, as it involves averages over the equilibrium distribution $P(\mathbf{v}|\mathbf{W},\mathbf{b}_v,\mathbf{b}_h)$. The parameter update for an RBM using CD learning is then given by

$$\Delta\theta \propto \left\langle \frac{\partial E_{RBM}}{\partial\theta}\right\rangle_0 - \left\langle \frac{\partial E_{RBM}}{\partial\theta}\right\rangle_n,$$

where the $<>_n$ denotes an average over the distribution P_n of the hidden and visible variables after n MCMC steps. The weight updates then become

$$\Delta\mathbf{W}_{i,j} \propto \frac{1}{\sigma^2}\langle v_i h_j\rangle_0 - \frac{1}{\sigma^2}\langle v_i h_j\rangle_n.$$

In general, $n=1$ already gives good results (Hinton and Salakhutdinov, 2006).

Autoencoders are deterministic models with two weight matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ representing the flow of data from the visible-to-hidden and hidden-to-visible layers, respectively (see Fig. 1B). AEs are trained to perform optimal reconstruction of the visible layer, often by minimizing the mean-squared error (MSE) in a reconstruction task. This is usually evaluated as follows: Given an activation pattern in the visible layer $\mathbf{v}$, we evaluate the activation of the hidden layer by $\mathbf{h} = sigm(\mathbf{v}^\top\mathbf{W}_1 + \mathbf{b}_h)$, where we will denote the bias in the hidden layer by $\mathbf{b}_h$. These activations are then propagated back to the visible layer through $\hat{v} = sigm(\mathbf{h}^\top\mathbf{W}_2 + \mathbf{b}_v)$ and the weights $\mathbf{W}_1$ and $\mathbf{W}_2$ are trained to minimize the distance measure between the original and reconstructed visible layers. Therefore, given a set of image samples $\{\mathbf{v}^d\}$ we can define the cost function. For example, using the squared Euclidean distance we have a cost function of

$$\mathcal{L}(\mathbf{W}_1,\mathbf{W}_2,\mathbf{b}_v,\mathbf{b}_h|\{\mathbf{v}^d\}) = \sum_d \|\mathbf{v}^d - \hat{\mathbf{v}}^d\|^2.$$

The weights can then be learned through stochastic gradient descent on the cost function. Autoencoders often yield better representations when trained on corrupted versions of the original data, performing gradient descent on the distance to the uncorrupted data. This approach is called a denoising Autoencoder (dAE) (Vincent et al., 2010). Note that in the AE, the activations of all units are continuous and not binary, and in general take values between 0 and 1.

4.1.2. *Existing dynamic models of unsupervised learning*

To date, a number of RBM-based models have been proposed to capture the sequential structure in time series data. Two of these models, the Temporal Restricted Boltzmann Machine and the Conditional Restricted Boltzmann machine, are introduced below.

Temporal Restricted Boltzmann Machines (TRBM) (Sutskever and Hinton, 2007) are a temporal extension of the standard RBM whereby feed-forward connections are included from previous time steps between hidden layers, from visible to hidden layers and from visible to visible layers (see Fig. 1D). Learning is conducted in the same manner as a normal RBM using contrastive divergence and it has been shown that such a model can be used to learn non-linear system evolutions such as the dynamics of a ball bouncing in a box (Sutskever and Hinton, 2007). A more restricted version of this model, discussed in Sutskever et al. (2008), can be seen in Fig. 1D and only contains temporal connections between the hidden layers. We will restrict ourselves to this model architecture in this paper.

Similarly to our notation for the RBM, we will write the visible layer variables as $\mathbf{v}^0,\ldots,\mathbf{v}^T$ and the hidden layer variables as $\mathbf{h}^0,\ldots,\mathbf{h}^T$. More precisely, $\mathbf{v}^T$ is the visible activation at the current time t and $\mathbf{v}^i$ is the visible activation at time $t-(T-i)$. The energy of the model for a given configuration of $\mathcal{V}=\{\mathbf{v}^0,\ldots,\mathbf{v}^T\}$ and $\mathcal{H}=\{\mathbf{h}^0,\ldots,\mathbf{h}^T\}$ is given by

$$E(\mathcal{H},\mathcal{V}|\mathcal{W}) = \sum_{t=0}^{T} E_{RBM}(\mathbf{h}^t,\mathbf{v}^t|\mathbf{W},\mathbf{b}) - \sum_{t=1}^{M}(\mathbf{h}^T)^\top\mathbf{W}_{T-t}\mathbf{h}^t, \quad (1)$$

where we have used $\mathcal{W}=\{\mathbf{W},\mathbf{W}_1,\ldots,\mathbf{W}_M\}$, where $\mathbf{W}$ are the static weights and $\mathbf{W}_1,\mathbf{W}_2,\ldots\mathbf{W}_M$ are the delayed weights for the temporally delayed hidden layers $\mathbf{h}_{T-1},\mathbf{h}_{T-2},\ldots,\mathbf{h}_0$ (see Fig. 1D). Note that, unlike the simple RBM, in the TRBM, the posterior distribution of any unit in the hidden layer conditioned on the visible layer is not independent of other hidden units, due to the connection between the delayed RBMs. This makes it harder to train the TRBM, as sampling from the hidden layer requires Gibbs sampling until the system has relaxed to its equilibrium distribution. This has led researcher to consider other types of probabilistic models for dynamic data.

Conditional Restricted Boltzmann Machines (CRBM) as described in Taylor et al. (2007) contain no temporal connections from the hidden layer but include connections from the visible layer at previous time steps to the current hidden and visible layers. The model architecture can be seen in Fig. 1C. In the CRBM, the past nodes are conditioned on, serving as a trial-specific bias. These units are shown in orange in Fig. 1C. Again, learning with this architecture requires only a small change to the energy function of the RBM and can be achieved through contrastive divergence. The CRBM is possibly the most successful of the Temporal RBM models to date and has been shown to both model and generate data from complex dynamical systems such as human motion capture data and video textures (Taylor, 2009).

4.1.3. *Temporal autoencoding training for TRBMs*

Much of the motivation for this work is to gain insight into the typical evolution of learned hidden layer features or RFs present in natural movie stimuli. With the existing CRBM this

Please cite this article as: Häusler, C., et al., Natural image sequences constrain dynamic receptive fields and imply a sparse code. Brain Research (2013), http://dx.doi.org/10.1016/j.brainres.2013.07.056

is not possible as it is unable to explicitly model the evolution of hidden features without resorting to a deep network architecture. Sparse coding models, as proposed by Cadieu and Olshausen (2008) overcome this restriction by learning complex filters, allowing for phase dynamics by multiplying the filters by complex weights whose dynamics are governed by phase variables. However, the evolution of the filters is indirectly modelled by the phase variables, not allowing for a direct biological interpretation.

The TRBM, in comparison, provides an explicit representation of the evolution of hidden features but, as we show, can be difficult to train using the standard algorithm. While this model does not have a direct biological influence, its artificial neural network structure allows for a biological interpretation of its function and indeed, producing a spiking neural network implementation of this approach would make for interesting future research. Here, we present a new pre-training method for the TRBM called Temporal Autoencoding (aTRBM) that dramatically improves its performance in modelling temporal data.

Training procedure: The energy of the model is given by Eq. (1) and is essentially an *M*-th order autoregressive RBM which is usually trained by standard contrastive divergence (Sutskever and Hinton, 2007). Here we propose to train it with a novel approach, highlighting the temporal structure of the stimulus. A summary of the training method is described in Table 1. First, the individual RBM visible-to-hidden weights $\mathbf{W}$ are initialized through contrastive divergence learning with a sparsity constraint on static samples of the dataset. After that, to ensure that the weights representing the hidden-to-hidden connections ($\mathbf{W}_t$) encode the dynamic structure of the ensemble, we initialize them by pre-training in the fashion of a denoising Autoencoder as will be described in the next section. After the Temporal Autoencoding is completed, the whole model (both visible-to-hidden and hidden-to-hidden weights) is trained together using contrastive divergence (CD) training.

One can regard the weights $\mathbf{W}$ as a representation of the static patterns contained in the data and the $\mathbf{W}_t$ as representing the transformation undergone by these patterns over time in the data sequences. This allows us to separate the representation of *form* and *motion* in the case of natural image sequences, a desirable property that is frequently studied in natural movies (see Cadieu and Olshausen, 2012). Furthermore, it allows us to learn how these features should evolve along time to encode the structure of the movies well. In the same way as static filters learned in this way often resemble RFs in visual cortex, the temporal projections learned here could be compared to lateral connections and correlations between neurons in visual cortex.

Temporal Autoencoding: The idea behind many feature extraction methods such as the autoencoder (Vincent et al., 2010) and reconstruction ICA (Le et al., 2011) is to find an alternative encoding for a set of data that allows for a good reconstruction of the dataset. This is frequently combined with sparse priors on the encoder. We propose to use a similar framework for TRBMs based on filtering (see Crisan and Rozovskii, 2011) instead of reconstructing through the use of a denoising Autoencoder (dAE). The key difference between an AE and a dAE is that random noise is added to each training sample before it is presented to the network, but the training procedure still requires the dAE to reproduce the *original* training data, before the noise was added, thereby *denoising* the training data. The addition of noise forces the model to learn reliable and larger scale structure from the training data as local perturbations from the added noise will change each time a sample is presented and are therefore unreliable.

In the aTRBM, we leverage the concept of denoising by treating previous samples of a sequential dataset as *noisy* versions of the current time point that we are trying to reproduce. The use of the term *noise* here is somewhat of a misnomer, but is used to keep in line with terminology from dAE literature. In the aTRBM case, no noise is added to the training data, but the small changes that exist between consecutive frames of the dataset are conceptually considered to be *noise* in the terms that we want to remove these changes from previous samples to be able to correctly reproduce or predict the data at the current time point. We can therefore use a dAE approach to constrain the temporal weights. In this sense, we consider the activity of the time-lagged visible units as noisy observations of the systems state, and want to infer the current state of the system. To this end, we propose pre-training the hidden-to-hidden weights of the TRBM by minimizing the error in predicting the present data frame from the previous observations of the data. This is similar to the approximation suggested by Sutskever et al. (2008), where the distribution over the hidden states conditioned on the visible history is approximated by the filtering distribution. The training is done as follows. After training the weights $\mathbf{W}$ we consider the model to be a deterministic Multi-Layer Perceptron with continuous activation in the hidden layers. We then consider the *M* delayed visible layers as features and try to predict the current visible layer by projecting through the hidden layers. In essence, we are considering the model to be a feed-forward network, where the delayed visible layers would form the input layer, the delayed hidden layers would constitute the first hidden layer, the current hidden layer would be the second hidden layer and the current visible layer would be the output. We can then write the prediction of the network as $\hat{\mathbf{v}}_d^T(\mathbf{v}_d^0, \mathbf{v}_d^1, \ldots, \mathbf{v}_d^{T-1})$, where the d index runs over the data points. The exact format of this function is described in Algorithm 1. We therefore minimize the reconstruction error given by

$$\mathcal{L}(\mathcal{W}) = \sum_d \left\| \mathbf{v}_d^T - \hat{\mathbf{v}}^T(\mathbf{v}_d^0, \mathbf{v}_d^1, \ldots, \mathbf{v}_d^{T-1}) \right\|^2,$$

where the sum over d goes over the entire dataset. The pre-training is described fully in Algorithm 1.

We train the temporal weights $\mathbf{W}_i$ one delay at a time, minimizing the reconstruction error with respect to that temporal weight stochastically. Then the next delayed temporal weight is trained keeping all the previous ones constant. The learning rate η is set adaptively during training following the advice given in Hinton (2010).

Algorithm 1. Pre-training temporal weights through Autoencoding.

for each sequence of data frames $I(t-T), I(t-(T-1)) \ldots, I(t)$, we take

$\mathbf{v}^T = I(t), \ldots, \mathbf{v}^0 = I(t-T)$ and **do**

Please cite this article as: Häusler, C., et al., Natural image sequences constrain dynamic receptive fields and imply a sparse code. Brain Research (2013), http://dx.doi.org/10.1016/j.brainres.2013.07.056

for each sequence of data frames $I(t-T), I(t-(T-1))\ldots, I(t)$, we take

  for $d=1$ to M **do**

    for $i=1$ to d **do**

      $\mathbf{h}^{T-i} = sigm(\mathbf{W}\mathbf{v}^{T-i} + \mathbf{b}_h)$

    end for

    $\mathbf{h}^{T} = sigm\left(\sum_{j=1}^{d} \mathbf{W}_j \mathbf{h}^{T-j} + \mathbf{b}_h\right)$, $\hat{\mathbf{v}}^{T} = sigm(\mathbf{W}^{\top}\mathbf{h}^{T} + \mathbf{b}_v)$

    $\epsilon(\mathbf{v}^{T}, \hat{\mathbf{v}}^{T}) = |\mathbf{v}^{T} - \hat{\mathbf{v}}^{T}|^2$

    $\Delta\mathbf{W}_d = \eta\partial\epsilon/\partial\mathbf{W}_d$

  end for

end for

4.2. Model analysis

4.2.1. Sparseness index

To measure spatial and temporal sparseness we employ the sparseness index introduced by Willmore and Tolhurst (2001) as

$$S = 1 - \frac{(\Sigma|a|/n)^2}{\Sigma(a^2/n)} \tag{2}$$

where a is the neural activation and n is the total number of samples used in the calculation. To quantify sparseness of the hidden unit activation we stimulate the aTRBM model that was previously trained on the Holywood2 dataset (cf. Section 2.2) with a single video sequences of approx. 30 s length at a frame rate of 30 s (total 897 frames) and measure the activation $\mathbf{h}$ of all hidden units during each video frame. *Spatial sparseness* refers to the distribution of activation values across the neuron population and is identical to the notion of population sparseness (Willmore et al., 2011). To quantify spatial sparseness we employ S to the activation values $\mathbf{h}$ across all 400 units for each of the time frames separately, resulting in 897 values. We use the notion of *temporal sparseness* to capture the distribution of activation values across time during a dynamic stimulus scenario (Haider et al., 2010). High temporal sparseness of a particular unit indicates that this unit shows strong activation only during a small number of stimulus frames. Low temporal sparseness indicates a flat activation curve across time. Our definition of temporal sparseness can easily be related to the definition of lifetime sparseness (Haider et al., 2010) if we consider each video frame as an independent stimulus. However, natural videos do exhibit correlations over time and successive video frames are thus generally not independent. Moreover, the dynamic RF model learns additional time dependencies. We employ S to quantify the temporal sparseness across the 897 single frame activation values for each neuron separately, resulting in 400 single unit measures.

Temporal and spatial sparseness are compared for the cases of a static RF and a dynamic RF. The static RF is defined by looking at the response of the aTRBM when all temporal weights are set to 0. This is equivalent to training a standard RBM.

4.2.2. Cascade spike generation model

From the activation variable h of the hidden units in our aTRBM model we generated spike train realizations using a cascade point process model (Herz et al., 2006) as described in (Fig. 6C). For each hidden unit we recorded its activation h during presentation of a video input. This time-varying activation expresses a probability between 0 and 1 of being active in each video frame. We linearly interpolated the activation curve to achieve a time resolution of 20 times the video frame rate. We then used the activation curve as intensity function to simulate single neuron spike train realizations according to the non-homogeneous Poisson process (Tuckwell, 2005). This can be generalized to other rate-modulated renewal and non-renewal point process models (Nawrot et al., 2008; Farkhooi et al., 2011). The expectation value for the trial-to-trial variability of the spike count is determined by the point process stochasticity (Nawrot et al., 2008) and thus independent of the activating model. We estimated neural firing rate from a single hidden neuron across repeated simulation trials or from the population of all 400 hidden neurons in a single simulation trial using the Peri Stimulus Time Histogram (Perkel et al., 1967; Nawrot et al., 1999; Shimazaki and Shinomoto, 2007) with a bin width corresponding to a single frame of the video input sequence.

4.3. Benchmark evaluation – human motion dynamics

We assessed the aTRBM's ability to learn a good representation of multi-dimensional temporal sequences by applying it to the 49 dimensional human motion capture data described by Taylor et al. (2007) and, using this as a benchmark, compared the performance to a TRBM without our pretraining method and Graham Taylor's example CRBM implementation.[2] All three models were implemented using Theano (Bergstra et al., 2010), have a temporal dependence of 6 frames (as in Taylor et al., 2007) and were trained using minibatches of 100 samples for 500 epochs.[3] The training time for all three models was approximately equal. Training was performed on the first 2000 samples of the dataset after which the models were presented with 1000 snippets of the data not included in the training set and required to generate the next frame in the sequence. For all three models, the visible-to-hidden connections were initialized with contrastive divergence on static snapshots of the data. For the TRBM we then proceeded to train all the weights of the model through contrastive divergence, whereas in the aTRBM case we initialized the weights through temporal autoencoding as described in Algorithm 1, before training the whole model with CD. The CRBM was also trained using contrastive divergence. In addition, we created a deterministic model which has the same structure as the aTRBM but was trained using only the first two training steps listed in Table 1 which we will refer to as an Autoencoded Multi-Layer Perceptron (AE/MLP).

[2]CRBM implementation available at https://gist.github.com/2505670.

[3]For the standard TRBM, training epochs were broken up into 100 static pretraining and 400 epochs for all the temporal weights together. For the aTRBM, training epochs were broken up into 100 static pretraining, 50 Autoencoding epochs per delay and 100 epochs for all the temporal weights together.

Please cite this article as: Häusler, C., et al., Natural image sequences constrain dynamic receptive fields and imply a sparse code. Brain Research (2013), http://dx.doi.org/10.1016/j.brainres.2013.07.056

Data generation in the aTRBM is done by taking a sample from the hidden layers at $t-6$ through $t-1$ and then Gibbs sampling from the RBM at time t while keeping the others fixed as biases. This is the filtering approximation from Sutskever et al. (2008). The visible layer at time t is initialized with noise and we sample for 30 Gibbs steps from the model. Data generation from the AE/MLP is done deterministically whereby the visible layers at $t-6$ through $t-1$ are set by the data and the activation is the propagated through to the visible layer at t for the sample prediction. We are interested in the performance of the AE/MLP to determine whether or not their is an advantage to the stochasticity of the RBM models in this prediction task. To this end, we also tested the deterministic performance of the three RBM models discussed here but the results were much poorer than those where the model generated data stochastically.

The results of a single trial prediction for four random dimensions of the dataset and the mean squared error (MSE) of the RBM model predictions over 100 repetitions for all 49 dimensions of the task can be seen in can be seen in Fig. 7. While the aTRBM is able to significantly outperform both the standard TRBM and CRBM models in this task during single trial prediction (3 leftmost columns), the deterministic AE/MLP model (middle column) predicts with an even lower error rate. In the 3 rightmost columns, we produce 50 single trial predictions per model type and take their mean as the prediction for the next frame in order to see if averaging over trials reduces the inherent variance of a single trial prediction. The performance of the CRBM and the aTRBM improve markedly and the aTRBM outperforms all other models. It should be noted that this process is not the same as taking the mean activation of the model (ie. a deterministic pass through the model with no sampling) which severely under performs the results shown here. Instead, averaging over multiple stochastic samples of the model proves to be advantageous in creating a low error estimate of the next frame. These results show not only the advantage of the aTRBM over the CRBM in this task, but also that of the stochastic models over the deterministic AE/MLP. Although single trial predictions from the aTRBM are not quite as accurate as those of the AE/MLP, the aTRBM is able to generate unique predictions stochastically at each trial, something the deterministic AE/MLP is not able to achieve. If one is interested purely in minimizing the MSE of the prediction, one can still use the aTRBM to generate and average over multiple trials which reduces the MSE and out performs the AE/MLP.

Acknowledgements

We thank Manfred Opper and Björn Kampa for helpful discussions. We also thank the reviewers of this manuscript for their constructive criticisms that led us to advance and refine this research. The work of Chris Häusler and Alex Susemihl was supported by the DFG Research Training Group Sensory Computation in Neural Systems (GRK 1589/1). The contribution of M.N. was funded by the German Federal Ministry of Education and Research within the Bernstein Focus Neuronal Basis of Learning (Grant no. 01GQ0941).

REFERENCES

Assisi, C., Stopfer, M., Laurent, G., Bazhenov, M., 2007. Adaptive regulation of sparseness by feedforward inhibition. Nat. Neurosci. 10 (9), 1176–1184.

Bell, A.J., Sejnowski, T.J., 1997. The independent components of natural scenes are edge filters. Vision Res. 37 (23), 3327–3338.

Benda, J., Herz, A.V., 2003. A universal model for spike-frequency adaptation. Neural Comput. 15 (11), 2523–2564.

Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H., 2007. Greedy layer-wise training of deep networks. Adv. Neural Inf. Process. Syst. 19, 153.

Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., Bengio, Y., 2010. Theano: a CPU and GPU math expression compiler. In: Proceedings of the Python for Scientific Computing Conference (SciPy). Oral Presentation.

Bosking, W., Zhang, Y., Schofield, B., Fitzpatrick, D., 1997. Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. J. Neurosci. 17 (6), 2112–2127.

Brill, M.F., Rosenbaum, T., Reus, I., Kleineidam, C.J., Nawrot, M.P., Rössler, W., 2013. Parallel processing via a dual olfactory pathway in the honeybee. J. Neurosci. 33 (6), 2443–2456.

Cadieu, C., Olshausen, B., 2012. Learning intermediate-level representations of form and motion from natural movies. Neural Comput., 1–40.

Cadieu, C.F., Olshausen, B.A., 2008. Learning transformational invariants from natural movies. In: Advances in Neural Information Processing Systems 21, pp. 1–8.

Carlson, N.L., Ming, V.L., DeWeese, M.R., 2012. Sparse codes for speech predict spectrotemporal receptive fields in the inferior colliculus. PLoS Comput. Biol. 8 (7), e1002594.

Caron, S.J., Ruta, V., Abbott, L., Axel, R., 2013. Random convergence of olfactory inputs in the drosophila mushroom body. Nature 497, 113–117.

Carreira-Perpinan, M., Hinton, G., 2005. On contrastive divergence learning. In: Artificial Intelligence and Statistics, vol. 2005, p. 17.

Cassenaer, S., Laurent, G., 2012. Conditional modulation of spike-timing-dependent plasticity for olfactory learning. Nature 482 (7383), 47–52.

Chacron, M.J., Longtin, A., Maler, L., 2001. Negative interspike interval correlations increase the neuronal capacity for encoding time-dependent stimuli. J. Neurosci. 21 (14), 5328–5343.

Chacron, M.J., Maler, L., Bastian, J., 2005. Electroreceptor neuron dynamics shape information transmission. Nat. Neurosci. 8 (5), 673–678.

Chen, C., Read, H., Escabí, M., 2012. Precise feature based time scales and frequency decorrelation lead to a sparse auditory code. J. Neurosci. 32 (25), 8454–8468.

Coppola, D., White, L., Fitzpatrick, D., Purves, D., 1998. Unequal representation of cardinal and oblique contours in ferret visual cortex. Proc. Natl. Acad. Sci. 95 (5), 2621–2623.

Crisan, D., Rozovskii, B., 2011. The Oxford Handbook of Nonlinear Filtering. Oxford University Press, Oxford.

Dan, Y., Atick, J., Reid, R., 1996. Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. J. Neurosci. 16 (10), 3351–3362.

Farkhooi, F., Froese, A., Muller, E., Menzel, R., Nawrot, M.P., 2012. Cellular adaptation accounts for the sparse and reliable sensory stimulus representation. http://arxiv.org/abs/1210.7165.

Farkhooi, F., Muller, E., Nawrot, M.P., 2011. Adaptation reduces variability of the neuronal population code. Phys. Rev. E 83 (5), 050905.

Please cite this article as: Häusler, C., et al., Natural image sequences constrain dynamic receptive fields and imply a sparse code. Brain Research (2013), http://dx.doi.org/10.1016/j.brainres.2013.07.056

Farkhooi, F., Strube-Bloss, M.F., Nawrot, M.P., 2009. Serial correlation in neural spike trains: experimental evidence, stochastic modeling, and single neuron variability. Phys. Rev. E 79 (2), 021905.

Field, D., Hayes, A., 2004. Contour integration and the lateral connections of v1 neurons. Visual Neurosci. 2, 1069–1079.

Friedrich, R.W., Laurent, G., 2004. Dynamics of olfactory bulb input and output activity during odor stimulation in zebrafish. J. Neurophysiol. 91 (6), 2658–2669.

Georgopoulos, A.P., Kalaska, J.F., Caminiti, R., Massey, J.T., 1982. On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. J. Neurosci. 2 (11), 1527–1537.

Haider, B., Krause, M.R., Duque, A., Yu, Y., Touryan, J., Mazer, J.A., McCormick, D.A., 2010. Synaptic and network mechanisms of sparse and reliable visual cortical activity during nonclassical receptive field stimulation. Neuron 65 (1), 107.

Herikstad, R., Baker, J., Lachaux, J.-P., Gray, C.M., Yen, S.-C., 2011. Natural movies evoke spike trains with low spike time variability in cat primary visual cortex. J. Neurosci. 31 (44), 15844–15860.

Herz, A.V., Gollisch, T., Machens, C.K., Jaeger, D., 2006. Modeling single-neuron dynamics and computations: a balance of detail and abstraction. Science 314 (5796), 80–85.

Hinton, G., 2010. A practical guide to training restricted Boltzmann machines. Momentum 9, 1.

Hinton, G., Salakhutdinov, R., 2006. Reducing the dimensionality of data with neural networks. Science 313 (5786), 504–507.

Hinton, G., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2012. Improving neural networks by preventing co-adaptation of feature detectors. Arxiv preprint arXiv:1207.0580.

Honegger, K.S., Campbell, R.A., Turner, G.C., 2011. Cellular-resolution population imaging reveals robust sparse coding in the Drosophila mushroom body. J. Neurosci. 31 (33), 11772–11785.

Hromádka, T., DeWeese, M., Zador, A., 2008. Sparse representation of sounds in the unanesthetized auditory cortex. PLoS Biol. 6 (1), e16.

Huerta, R., 2013. Learning pattern recognition and decision making in the insect brain. In: American Institute of Physics Conference Series, vol. 1510, pp. 101–119.

Huerta, R., Nowotny, T., 2009. Fast and robust learning by reinforcement signals: explorations in the insect brain. Neural Comput. 21 (8), 2123–2151.

Huerta, R., Nowotny, T., García-Sanchez, M., Abarbanel, H., Rabinovich, M., 2004. Learning classification in the olfactory system of insects. Neural Comput. 16 (8), 1601–1640.

Ito, I., Ong, R.C.-y., Raman, B., Stopfer, M., 2008. Sparse odor representation and olfactory learning. Nat. Neurosci. 11 (10), 1177–1184.

Jadhav, S., Wolfe, J., Feldman, D., 2009. Sparse temporal coding of elementary tactile features during active whisker sensation. Nat. Neurosci. 12 (6), 792–800.

Jortner, R.A., Farivar, S.S., Laurent, G., 2007. A simple connectivity scheme for sparse coding in an olfactory system. J. Neurosci. 27 (7), 1659–1669.

Knoblauch, A., Palm, G., et al., (2001). Pattern separation and synchronization in spiking associative memories and visual areas. Neural Networks: Off. J. Int. Neural Network Soc. 14 (6–7), 763.

Krofczik, S., Menzel, R., Nawrot, M.P., 2008. Rapid odor processing in the honeybee antennal lobe network. Front. Comput. Neurosci., 2.

Larochelle, H., Bengio, Y., 2008. Classification using discriminative restricted Boltzmann machines. In: Proceedings of the 25th International Conference on Machine Learning, ACM, pp. 536–543.

Laughlin, S.B., et al., (2001). Energy as a constraint on the coding and processing of sensory information. Curr. Opin. Neurobiol. 11 (4), 475–480.

Le, Q., Karpenko, A., Ngiam, J., Ng, A., 2011. Ica with reconstruction cost for efficient overcomplete feature learning. In: NIPS.

Lee, H., Ekanadham, C., Ng, A., 2008. Sparse deep belief net model for visual area v2. Adv. Neural Inf. Process. Syst. 20, 873–880.

Lee, H., Largman, Y., Pham, P., Ng, A., 2009. Unsupervised feature learning for audio classification using convolutional deep belief networks. Adv. Neural Inf. Process. Syst. 22, 1096–1104.

Lennie, P., 2003. The cost of cortical computation. Curr. Biol. 13 (6), 493–497.

Levy, W.B., Baxter, R.A., 1996. Energy efficient neural codes. Neural Comput. 8 (3), 531–543.

Maldonado, P., Babul, C., Singer, W., Rodriguez, E., Berger, D., Grün, S., 2008. Synchronization of neuronal responses in primary visual cortex of monkeys viewing natural images. J. Neurophysiol. 100 (3), 1523–1532.

Marszalek, M., Laptev, I., Schmid, C., 2009. Actions in context. In: IEEE Conference on Computer Vision and Pattern Recognition.

Martin, K.A., Schröder, S., 2013. Functional heterogeneity in neighboring neurons of cat primary visual cortex in response to both artificial and natural stimuli. J. Neurosci. 33 (17), 7325–7344.

Mohamed, A., Sainath, T., Dahl, G., Ramabhadran, B., Hinton, G., Picheny, M., 2011. Deep belief networks using discriminative features for phone recognition. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 5060–5063.

Nawrot, M., Aertsen, A., Rotter, S., 1999. Single-trial estimation of neuronal firing rates: from single-neuron spike trains to population activity. J. Neurosci. Methods 94 (1), 81–92.

Nawrot, M.P., 2010. Analysis and interpretation of interval and count variability in neural spike trains. In: Analysis of Parallel Spike Trains, Springer, pp. 37–58.

Nawrot, M.P., 2012. Dynamics of sensory processing in the dual olfactory pathway of the honeybee. Apidologie 43 (3), 269–291.

Nawrot, M.P., Boucsein, C., Rodriguez-Molina, V., Aertsen, A., Grün, S., Rotter, S., 2007. Serial interval statistics of spontaneous activity in cortical neurons *in vivo* and *in vitro*. Neurocomputing 70 (10), 1717–1722.

Nawrot, M.P., Boucsein, C., Rodriguez Molina, V., Riehle, A., Aertsen, A., Rotter, S., 2008. Measurement of variability dynamics in cortical spike trains. J. Neurosci. Methods 169 (2), 374–390.

Olshausen, B., et al., (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381 (6583), 607–609.

Olshausen, B.A., Field, D.J., et al., (2004). Sparse coding of sensory inputs. Curr. Opin. Neurobiol. 14 (4), 481–487.

Palm, G., 1980. On associative memory. Biol. Cybern. 36 (1), 19–31.

Papadopoulou, M., Cassenaer, S., Nowotny, T., Laurent, G., 2011. Normalization for sparse encoding of odors by a wide-field interneuron. Science 332 (6030), 721–725.

Perez-Orive, J., Mazor, O., Turner, G.C., Cassenaer, S., Wilson, R.I., Laurent, G., 2002. Oscillations and sparsening of odor representations in the mushroom body. Science 297 (5580), 359–365.

Perkel, D.H., Gerstein, G.L., Moore, G.P., 1967. Neuronal spike trains and stochastic point processes: I. the single spike train. Biophys. J. 7 (4), 391–418.

Pfeil, T., Grübl, A., Jeltsch, S., Müller, E., Müller, P., Petrovici, M.A., Schmuker, M., Brüderle, D., Schemmel, J., Meier, K., 2013. Six networks on a universal neuromorphic computing substrate. Front. Neurosci., 7.

Please cite this article as: Häusler, C., et al., Natural image sequences constrain dynamic receptive fields and imply a sparse code. Brain Research (2013), http://dx.doi.org/10.1016/j.brainres.2013.07.056

Poo, C., Isaacson, J.S., 2009. Odor representations in olfactory cortex: sparse coding, global inhibition and oscillations. Neuron 62 (6), 850.

Reich, D.S., Mechler, F., Victor, J.D., 2001. Independent and redundant information in nearby cortical neurons. Science 294 (5551), 2566–2568.

Reinagel, P., 2001. How do visual neurons respond in the real world?. Curr. Opin. Neurobiol. 11 (4), 437–442.

Reinagel, P., Reid, R., 2002. Precise firing events are conserved across neurons. J. Neurosci. 22 (16), 6837–6841.

Rickert, J., Riehle, A., Aertsen, A., Rotter, S., Nawrot, M.P., 2009. Dynamic encoding of movement direction in motor cortical neurons. J. Neurosci. 29 (44), 13870–13882.

Riehle, A., Grün, S., Diesmann, M., Aertsen, A., 1997. Spike synchronization and rate modulation differentially involved in motor cortical function. Science 278 (5345), 1950–1953.

Saxe, A., Bhand, M., Mudur, R., Suresh, B., Ng, A., 2011. Unsupervised learning models of primary cortical receptive fields and receptive field plasticity. In: Advances in Neural Information Processing Systems.

Shimazaki, H., Shinomoto, S., 2007. A method for selecting the bin size of a time histogram. Neural Comput. 19 (6), 1503–1527.

Simoncelli, E., Olshausen, B., 2001. Natural image statistics and neural representation. Annu. Rev. Neurosci. 24 (1), 1193–1216.

Strube-Bloss, M.F., Nawrot, M.P., Menzel, R., 2011. Mushroom body output neurons encode odor–reward associations. J. Neurosci. 31 (8), 3129–3140.

Sutskever, I., Hinton, G., 2007. Learning multilevel distributed representations for high-dimensional sequences. In: Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, pp. 544–551.

Sutskever, I., Hinton, G., Taylor, G., 2008. The recurrent temporal restricted Boltzmann machine. Adv. Neural Inf. Process. Syst., 21.

Taylor, G., 2009. Composable, Distributed-state Models for High-dimensional Time Series. Ph.D. Thesis.

Taylor, G., Hinton, G., Roweis, S., 2007. Modeling human motion using binary latent variables. Adv. Neural Inf. Process. Syst. 19, 1345.

Tuckwell, H.C., 2005. Introduction to Theoretical Neurobiology: vol. 2, Nonlinear and Stochastic Theories, vol. 8. Cambridge University Press.

van Hateren, J.H., Ruderman, D.L., 1998. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. Proc. R. Soc. London Ser. B: Biol. Sci. 265 (1412), 2315–2320.

Van Hooser, S., 2007. Similarity and diversity in visual cortex: is there a unifying theory of cortical computation?. The Neuroscientist 13 (6), 639–656.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P., 2010. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. 11, 3371–3408.

Vinje, W., Gallant, J., 2000. Sparse coding and decorrelation in primary visual cortex during natural vision. Science 287 (5456), 1273–1276.

Wang, G., Ding, S., Yunokuchi, K., 2003. Difference in the representation of cardinal and oblique contours in cat visual cortex. Neurosci. Lett. 338 (1), 77–81.

Willmore, B., Tolhurst, D.J., 2001. Characterizing the sparseness of neural codes. Network: Comput. Neural Syst. 12 (3), 255–270.

Willmore, B.D., Mazer, J.A., Gallant, J.L., 2011. Sparse coding in striate and extrastriate visual cortex. J. Neurophysiol. 105 (6), 2907–2919.

Wilson, R.I., Turner, G.C., Laurent, G., 2004. Transformation of olfactory representations in the Drosophila antennal lobe. Sci. Signal. 303 (5656), 366.

Wolfe, J., Houweling, A., Brecht, M., et al., (2010). Sparse and powerful cortical spikes. Curr. Opin. Neurobiol. 20 (3), 306.

Yen, S., Baker, J., Gray, C., 2007. Heterogeneity in the responses of adjacent neurons to natural stimuli in cat striate cortex. J. Neurophysiol. 97 (2), 1326–1341.

Zetzsche, C., Nuding, U., 2005. Nonlinear and higher-order approaches to the encoding of natural scenes. Network: Comput. Neural Syst. 16 (2–3), 191–221.

Please cite this article as: Häusler, C., et al., Natural image sequences constrain dynamic receptive fields and imply a sparse code. Brain Research (2013), http://dx.doi.org/10.1016/j.brainres.2013.07.056