Spatiotemporal Receptive Fields Maximizing Temporal Coherence in Natural Image Sequences

Jarmo Hurri [a] and Jaakko Väyrynen [a] and Aapo Hyvärinen [b,a]

[a] *Neural Networks Research Centre*

Helsinki University of Technology, P.O.Box 9800, 02015 HUT, Finland

[b] *Helsinki Institute for Information Technology / BRU*

Department of Computer Science

University of Helsinki, P.O.Box 26, 00014 UH, Finland

Abstract

The relationship between the structure and functionality of the visual system and the properties of natural visual stimuli is an active research topic in computational visual neuroscience. It has previously been shown that maximization of temporal coherence of activity levels is one computational principle which leads to the emergence of simple-cell-like spatial receptive fields from natural image sequences. In this paper we extend previous results by examining the case of spatiotemporal receptive fields. We show that application of the same principle of temporal coherence in the spatiotemporal case yields receptive fields which are not only localized, oriented and multiscale, as in the spatial case, but also share some important temporal characteristics with spatiotemporal simple-cell receptive fields. Quantitative measurements of the properties of the resulting receptive fields are also provided, and these are compared against similar results obtained with independent component analysis.

Preprint submitted to Elsevier Science *20 October 2003*

Key words: natural image sequences, primary visual cortex, temporal coherence

1 Introduction

Within visual computational neuroscience, there is an area of research where it is assumed that the statistical properties of natural visual stimuli have influenced the functional properties of cells in the early visual system (for a review, see [1]). In this field, studies concerning the primary visual cortex typically employ linear filters as models of the classical receptive fields (CRFs) of simple cells. The case of spatial receptive-field models has been studied for the most prominent computational theories: temporal coherence and independent component analysis (ICA) / sparse coding (e.g., [2,3]). Some studies employing spatiotemporal models in ICA and sparse coding have also been published [4,5]. However, in the case of temporal coherence, no studies employing spatiotemporal CRFs have been reported. In this paper, we use temporal coherence to learn spatiotemporal CRFs from natural image sequence data. There are three main contributions in this paper: the qualitative and quantitative description of temporally coherent spatiotemporal CRFs, the comparison of results obtained with temporal coherence and independent component analysis against each other, and the comparison of the results obtained with these methods against recent, comprehensive physiological data [6].

Email addresses: jarmo.hurri@hut.fi (Jarmo Hurri), jaakko.j.vayrynen@hut.fi (Jaakko Väyrynen), aapo.hyvarinen@hut.fi (Aapo Hyvärinen).

2 Temporal coherence of activity levels

The core idea of temporal coherence is that the neural representation changes as little as possible over time, while still preserving (almost) all of the information about the input data. It has previously been shown that maximization of *temporal coherence of activity levels* is one computational principle which leads to the emergence of simple-cell-like spatial CRFs from natural image sequences [3]. In this paper we study the case of spatiotemporal CRFs, presented here in a matrix-vector formulation. A *vectorization* of spatiotemporal image sequence samples can be done by scanning the frames of an image sequence one by one column-wise into a vector. Let a vectorized sequence of frames, taken from natural video at time t, be denoted $\mathbf{x}(t)$. Let $\mathbf{y}(t)$ represent the outputs of K simple cells: $\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t)$. Here, $\mathbf{W} = [\mathbf{w}_1 \cdots \mathbf{w}_K]^T$ denotes a matrix with all the CRFs as rows.

Temporal response strength correlation [3], the objective function, is defined by

$$f(\mathbf{W}) = \sum_{k=1}^{K} \mathrm{E}_t \left\{ g(y_k(t)) g(y_k(t - \Delta t)) \right\}, \tag{1}$$

where the nonlinearity g is strictly convex, even (rectifying), and differentiable (such as $g(x) = \ln \cosh x$), and Δt denotes a delay in time. A set of CRFs which has a large temporal response strength correlation is such that the same neurons *often respond strongly at consecutive time points*, outputting large (either positive or negative) values, thereby expressing temporal coherence of a population code. Additional constraints are used to keep the outputs of the neurons bounded and to keep the CRFs from converging to the same solution, and a gradient projection method can be used to maximize the resulting constrained optimization problem [3]. The initial value of $\mathbf{W}$ is

selected randomly. One standard way to interpret the results obtained with linear simple-cell models is to express the relationship between data $\mathbf{x}(t)$ and neural responses $\mathbf{y}(t)$ as a *generative model* [2,5]: $\mathbf{x}(t) = \mathbf{A}\mathbf{y}(t)$, where $\mathbf{A}$ is given by the inverse (or pseudoinverse) of $\mathbf{W}$. Each column of matrix $\mathbf{A}$ can be interpreted as the feature coded by the corresponding simple cell. Below we will use this interpretation when we present our results.

3 Data collection and preprocessing

The data used in the experiments was sampled from the database of natural image sequences described in [4]. The sampled data consisted of 120,000 image sequence blocks of size $11 \times 11 \times 9$ ($x \times y \times t$). Each sample of length 9 frames was divided into two partially overlapping samples of length 8; this yields $\mathbf{x}(t)$ and $\mathbf{x}(t-\Delta t)$. The sampling frequency of the data was 25 Hz, so Δt was 40 ms, and the durations of $\mathbf{x}(t)$ and $\mathbf{x}(t - \Delta t)$, and the spatiotemporal CRF, were 280 ms. Preprocessing consisted of removal of local average image intensity and dimensionality reduction by 50% to 484 using principal component analysis [7]. Dimensionality reduction reduces the effect of noise and aliasing artifacts, and decreases the computational complexity of the problem (the degree of dimensionality reduction applied here retains 95% of original signal energy).

4 Results and discussion

Some of the resulting spatiotemporal basis vectors of size $11 \times 11 \times 8$ (i.e., columns of $\mathbf{A}$) maximizing objective function (1) are shown in Fig. 1. As can be seen, the learned receptive fields share the primary spatial properties of

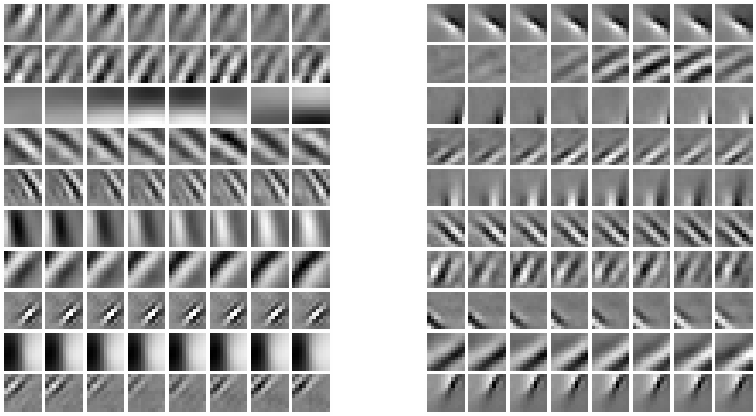

Fig. 1. A subset of 20 spatiotemporal receptive field models (columns of $\mathbf{A}$) obtained by maximizing temporal coherence of activity levels in natural image sequences (10 receptive fields in the image on the left and 10 on the right). Each of the 20 rows corresponds to one spatiotemporal receptive field model, and the frames in a row correspond to spatial snapshots of the spatiotemporal receptive field at consecutive time instances.

simple cells in that they are localized, oriented, and have multiple scales (see, e.g., [8]). In addition to these spatial properties, the receptive fields also have physiologically relevant qualitative temporal properties. For example, some of the receptive fields seem to be space-time separable [6], while others are inseparable.[1] Also, different space-time inseparable receptive fields seem to respond to different velocities. To obtain a corresponding set of ICA[2] results, we applied the symmetric fast fixed-point ICA algorithm [7] with nonlinearity $g(y) = \tanh y$ to the same data.

To assess the results quantitatively, we measured some important parameters from the two sets of CRFs. The results of these measurements are shown

[1] A space-time separable receptive field can be expressed as a product of a one-dimensional temporal profile and a two-dimensional spatial profile.

[2] Independent component analysis is a well-known method (see, e.g., [7]), so it will not be explained here.

in Fig. 2, with Figs. 2A–E containing spatial measurement information, and Figs. 2F–J containing temporal measurement information. The operational definitions of these quantities can be found in [6]. As a general observation, the histograms of measured parameters for temporal coherence and ICA are mostly similar. Therefore, in what follows we focus on the comparison against physiological measurements. These measurements, shown in Fig. 3, were made by DeAngelis et al. [6] from 91 simple cells. Because of the relatively low number of measurement points (91 cells), we will only consider the distributions qualitatively, which still turns out to produce some interesting results.

When the results obtained with temporal coherence and ICA are compared against the physiological measurements, we see similarities in the spatial measurements, and major differences in the temporal measurements. In the spatial measurements (Figs. 2A–E and 3A–E), the distributions have similar qualitative properties, although the number of subregions is substantially higher in the computational results. In the case of temporal measurements (Figs. 2F–J and 3F–J), the physiological measurements are strikingly different from the measurements of learned CRFs. In all other cases except direction selectivity, the histograms of the physiological measurements look almost completely different from the corresponding histograms obtained from the two models. The key measurement in understanding these differences is receptive field duration (Figs. 2F and 3F). Practically all of the CRFs which emerge from the two models span the whole time frame of the receptive field (see also Fig. 1). When the CRF has a practically constant magnitude over the whole time frame, the point where the maximum is reached is somewhat arbitrary (Fig. 2G). The differences in optimal temporal frequency (Figs. 2H and 3H) are probably also related to the lack of temporal change.

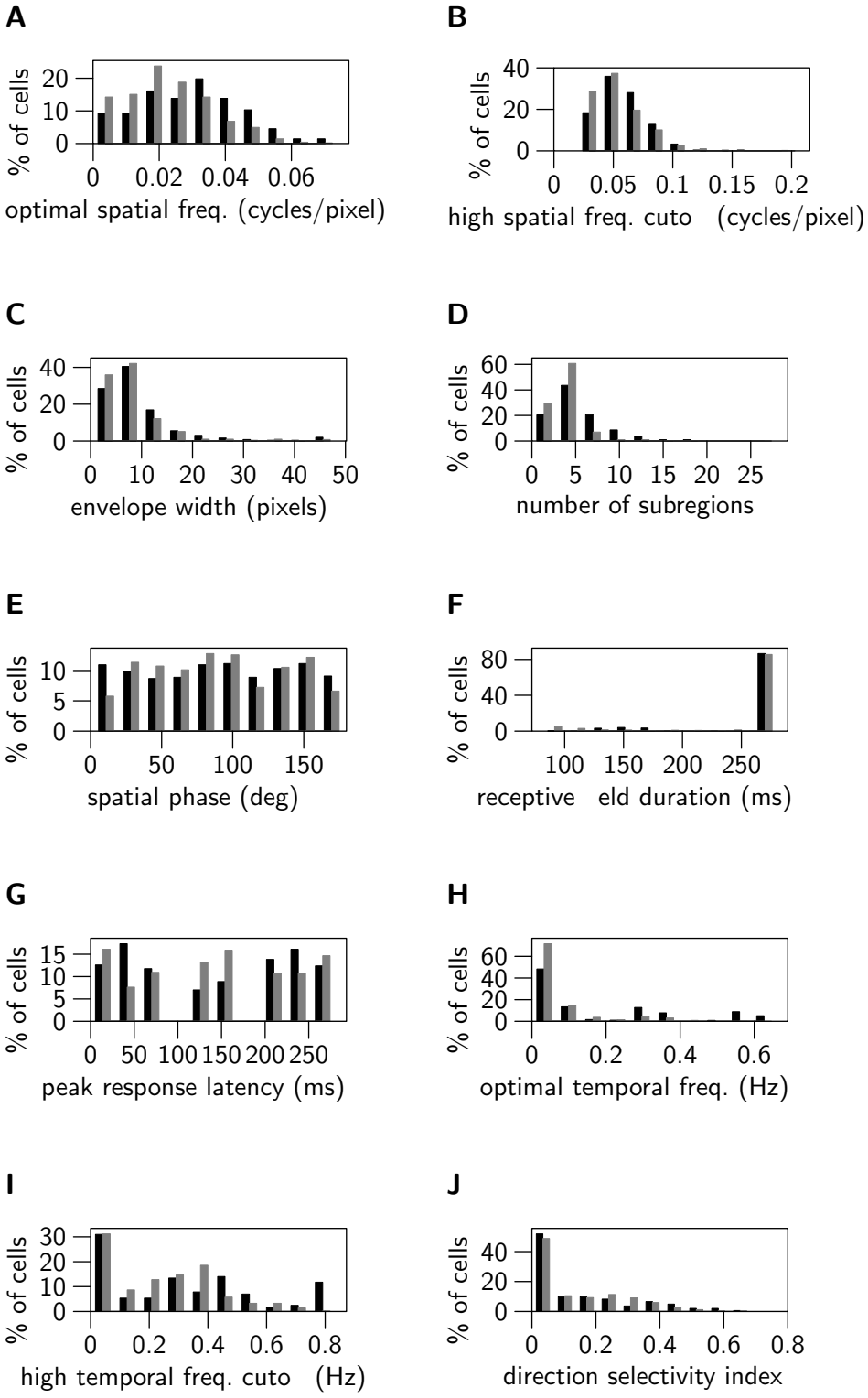

Fig. 2. Quantitative measurements of spatiotemporal classical receptive fields obtained with temporal coherence (black bars) and ICA (grey bars).

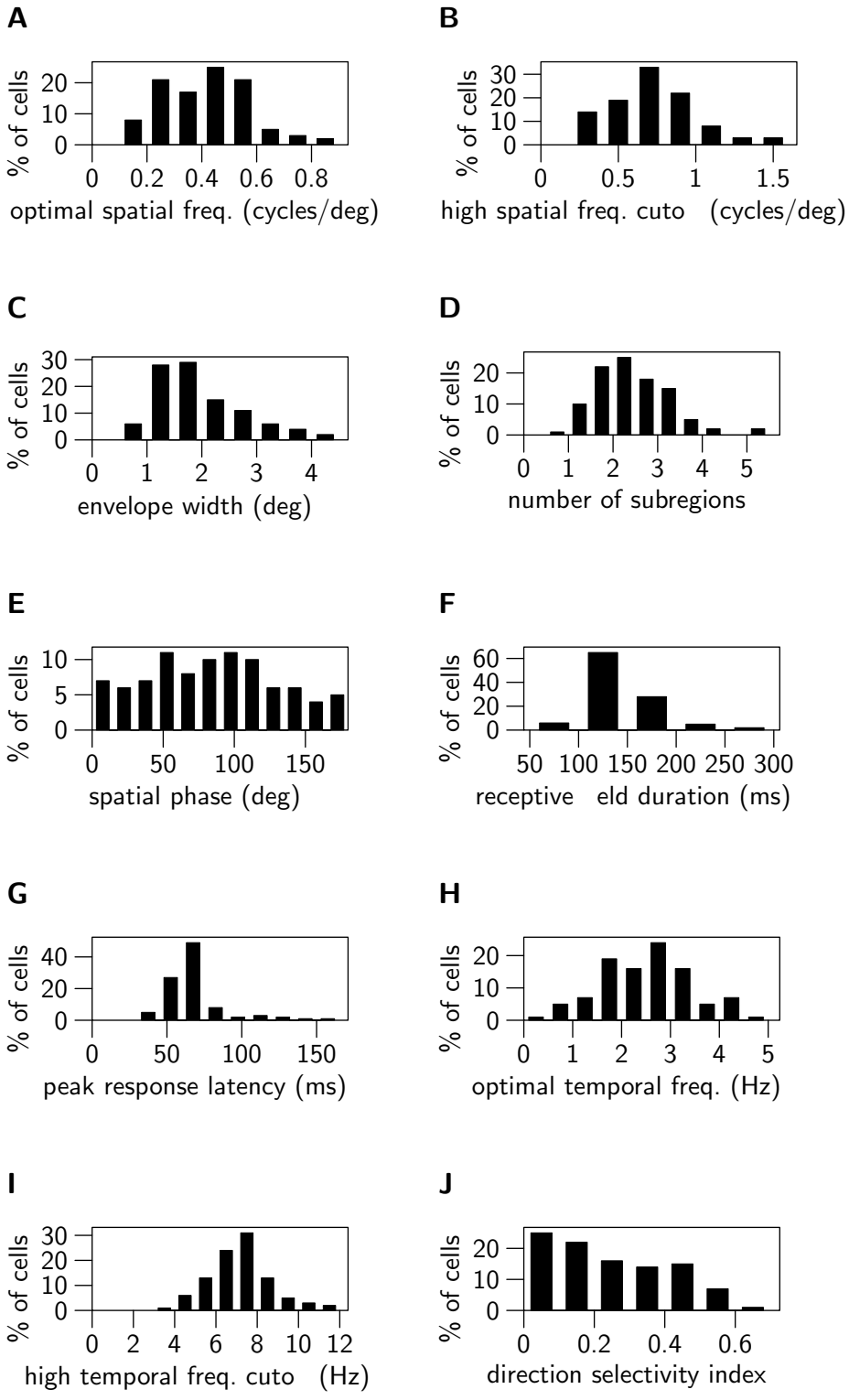

Fig. 3. For comparison, results reported by DeAngelis et al.: physiological spatiotemporal receptive field measurements made from adult cats [6]. Note that in subfigures A–C, x-axis units differ from Figs. 2A–C, and that in subfigures D and F–I, x-axis limits differ from Figs. 2D and 2F–I.

Currently we are unable to provide an adequate explanation to the lack of temporal localization in our results. In previous research, van Hateren and Ruderman have applied an ICA algorithm to learn spatiotemporal CRFs from natural image sequences [4]. In their results, the CRFs are more localized in time. In [4] different preprocessing methods and a slightly different ICA algorithm were used, and only a subset of a complete basis was computed, which might explain the differences in their ICA results and ours. We have performed an additional set of experiments to study the effects of different preprocessing methods and basis sizes. Although we have obtained some results with slightly better temporal localization, we have not been able to link the differences in preprocessing or basis size conclusively to the qualitative differences between our temporal coherence results and the ICA results in [4]. This difference needs to be studied in more detail in further research.

5 Conclusions

In this paper, we have applied both temporal coherence and independent component analysis to learn spatiotemporal receptive field models from natural image sequences. Our results show that the results obtained with these two methods have similar spatial and temporal quantitative properties. When compared with physiological measurements from cat simple cells, similarities between the learned CRFs and physiological measurements can be found in the spatial domain, while there are substantial differences in the temporal domain.

References

[1] Eero P. Simoncelli and Bruno A. Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24:1193–1216, 2001.

[2] Bruno A. Olshausen and David Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[3] Jarmo Hurri and Aapo Hyvärinen. Simple-cell-like receptive fields maximize temporal coherence in natural video. *Neural Computation*, 15(3):663–691, 2003.

[4] J. Hans van Hateren and Dan L. Ruderman. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proceedings of the Royal Society of London B*, 265(1412):2315–2320, 1998.

[5] Bruno A. Olshausen. Sparse coding of time-varying natural images. In Petteri Pajunen and Juha Karhunen, editors, *Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation*, pages 603–608, 2000.

[6] Gregory C. DeAngelis, Izumi Ohzawa, and Ralph D. Freeman. Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. I. General characteristics and postnatal development. *Journal of Neurophysiology*, 69(4):1091–1117, 1993.

[7] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

[8] Stephen E. Palmer. *Vision Science – Photons to Phenomenology*. The MIT Press, 1999.