Network: Comput. Neural Syst. **11** (2000) 191–210. Printed in the UK PII: S0954-898X(00)14814-6

Independent component analysis applied to feature extraction from colour and stereo images

Patrik O Hoyer†§ and Aapo Hyvärinen†‡

† Neural Networks Research Centre, Helsinki University of Technology, PO Box 5400, FIN-02015 HUT, Finland

‡ Department of Psychology, General Psychology Division, University of Helsinki, PO Box 13, FIN-00014 HUT, Finland

E-mail: patrik.hoyer@hut.fi

Received 24 January 2000

Abstract. Previous work has shown that independent component analysis (ICA) applied to feature extraction from natural image data yields features resembling Gabor functions and simple-cell receptive fields. This article considers the effects of including chromatic and stereo information. The inclusion of colour leads to features divided into separate red/green, blue/yellow, and bright/dark channels. Stereo image data, on the other hand, leads to binocular receptive fields which are tuned to various disparities. The similarities between these results and the observed properties of simple cells in the primary visual cortex are further evidence for the hypothesis that visual cortical neurons perform some type of redundancy reduction, which was one of the original motivations for ICA in the first place. In addition, ICA provides a principled method for feature extraction from colour and stereo images; such features could be used in image processing operations such as denoising and compression, as well as in pattern recognition.

1. Introduction

Ever since the classic experiments of Hubel and Wiesel (1962, 1968) there have been a large number of studies in which the receptive field properties of neurons in the primary visual cortex have been measured‖. The general consensus is that most receptive fields are localized in space and time, have band-pass characteristics in spatial and temporal frequency and are selective to some preferred orientation. In addition, many are selective to direction of movement, chromatic contrast, and/or binocular disparity (see, for example, DeValois *et al* 1982, DeAngelis *et al* 1993a, Livingstone and Hubel 1984, Barlow *et al* 1967).

Why do the neurons respond the way they do? The proposal by Barlow (1989) is that the neurons perform redundancy reduction, and make up a factorial code for the input data, i.e. a representation with independent components. Representing the data in this way would be useful for detecting new patterns, or 'suspicious coincidences'. Field (1994) has argued that oriented edge features constitute a sparse representation of the images. This means that for any one image, only a few of the features are needed to represent that particular image; and that over an ensemble of images a particular feature will seldom be significantly active.

§ Author to whom correspondence should be addressed.

‖ Most investigations have concerned the visual cortex of cats and monkeys. It is generally believed that the receptive fields of neurons in the human primary visual cortex are qualitatively similar.

0954-898X/00/030191+20$30.00

The possible benefits of sparse coding include increasing the signal-to-noise ratio and aiding in pattern recognition (Field 1994).

Recently, these theories have been tested experimentally. Olshausen and Field (1996, 1997) applied a sparseness-maximization network to input data consisting of image patches from natural images. Basically, one attempts to represent each image patch as a linear combination of 'basis' patches, such that the mixing coefficients are as sparse as possible. In other words,

$$\boldsymbol{x} = \mathbf{A}\boldsymbol{s} = \sum_{i=1}^{n} \boldsymbol{a}_i s_i \tag{1}$$

where we have denoted the input patch by $\boldsymbol{x}$, and the basis patches are the $\boldsymbol{a}_i$, the columns of $\mathbf{A}$ (see figure 1). One then optimizes the $\boldsymbol{a}_i$ such that for typical $\boldsymbol{x}$, most of the s_i will be close to zero and only a few will have significantly non-zero values. This led to features qualitatively similar to simple-cell receptive fields.

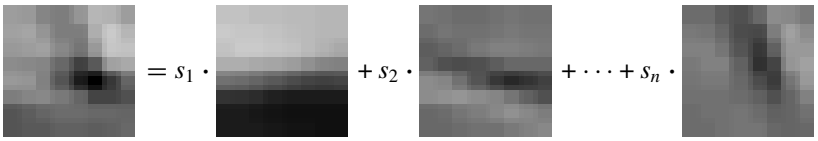

Figure 1. The linear image synthesis model. Each patch is represented as a linear combination of basis patches. In sparse coding, one attempts to find a representation such that the coefficients s_i are as 'sparse' as possible, meaning that for most image patches only a few of them are significantly active. In ICA, the purpose is to find a representation such that they are mutually as statistically independent as possible (cf equation (1)).

Subsequently, Bell and Sejnowski (1997) as well as Hurri *et al* (1997) applied independent component analysis (ICA) to similar data. In ICA, the decomposition is also linear as in (1), but now the purpose is to seek mutually independent components s_i. In other words, we seek a factorial code for the data. This leads to a local representation quite similar to that obtained through sparse coding. In fact, it has been shown (Olshausen and Field 1997) that the sparseness-maximization network and ICA are very closely related. Later, van Hateren and van der Schaaf (1998) quantitatively compared the filters learned by ICA to measurements of neural receptive fields, and found a good match for most parameters.

The extension of the above experiments into the spatio-temporal domain was performed by van Hateren and Ruderman (1998). Instead of considering only static image patches, they took sequential patches from video sequences to take into account temporal redundancy. Again, the found ICA decomposition seemed to fit the receptive field properties found in the cortex, giving features tuned to temporal as well as spatial frequencies.

This article reports on similar numerical experiments on two of the remaining dimensions of our visual input: chromatic contrast and stereopsis. We extract features from natural colour and stereo images using ICA, and compare the results to known receptive field properties of neurons found in the visual cortex. The paper is structured as follows. Section 2 discusses the ICA model and how it is usually applied to image data. In section 3, we report on experiments with colour images as input data, whereas section 4 is devoted to stereo image results. Section 5 states the connections to previous work while section 6 gives some conclusions.

2. ICA and image data

2.1. Preprocessing and estimation

As discussed in the introduction, in ICA (Jutten and Herault 1991, Comon 1994) we attempt to linearly transform the data to obtain statistically independent components. Recently, there has been a considerable amount of research on algorithms for performing ICA (Amari *et al* 1996, Bell and Sejnowski 1995, Cardoso and Laheld 1996, Cichocki and Unbehauen 1996, Hyvärinen and Oja 1997, Hyvärinen 1999a, Karhunen *et al* 1997, Oja 1997, Pajunen 1998). For a survey of research on ICA we refer the reader to Hyvärinen (1999c). Here, we will only give a brief introduction to ICA in the context of image data. For a more detailed discussion see, for example, Bell and Sejnowski (1997) and Hyvärinen *et al* (2000).

Assume that image patches of the type shown in figure 1 are represented as samples of a random vector $\boldsymbol{x}$. The goal then is to express the data by a linear generative model (1) where the stochastic sources (s_i) are as mutually independent as possible. This is what ICA algorithms attempt to perform. There are two quite standard preprocessing steps in ICA. First, the mean of the data is usually subtracted to centre the data on the origin. In other words,

$$\boldsymbol{x} := \boldsymbol{x} - E\{\boldsymbol{x}\}. \tag{2}$$

This does not alter the ICA model (1) except that we now have zero-mean sources: $E\{\boldsymbol{s}\} = 0$. The second step is to *whiten* the data. This means that we transform the data so the components are uncorrelated and have unit variance:

$$\boldsymbol{z} = \mathbf{V}\boldsymbol{x} \qquad \text{so that} \qquad E\{\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}\} = \mathbf{I} \tag{3}$$

where $\mathbf{V}$ is the whitening matrix and $\boldsymbol{z}$ the whitened data. This does not completely specify $\mathbf{V}$; in fact, if $\mathbf{V}$ is any solution then $\mathbf{WV}$ is also a whitening matrix for any orthogonal $\mathbf{W}$, as

$$E\{\mathbf{W}\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\} = \mathbf{W}E\{\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}\}\mathbf{W}^{\mathrm{T}} = \mathbf{W}\mathbf{W}^{\mathrm{T}} = \mathbf{I}. \tag{4}$$

For image data there are two commonly used and analytically available solutions (Bell and Sejnowski 1997). First, there is the symmetric whitening matrix $\mathbf{V}_{\mathrm{ZCA}} = E\{\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}^{-1/2}$. This is the local solution, where each filter whitens a local region of the input. Each filter is identical (neglecting border effects); basically a centre-on surround-off filter. It has been suggested that the centre-surround receptive fields of neurons in the retina and the lateral geniculate nucleus (LGN) perform something similar to symmetric whitening (Atick and Redlich 1993, Dan *et al* 1996).

Second, there is the principal component analysis (PCA) solution. Here, $\mathbf{V}_{\mathrm{PCA}} = \mathbf{D}^{-1/2}\mathbf{E}^{\mathrm{T}}$, where $\mathbf{EDE}^{\mathrm{T}} = E\{\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$ is the eigensystem of the correlation matrix of $\boldsymbol{x}$. In PCA, the filters (rows of $\mathbf{V}_{\mathrm{PCA}}$) are orthogonal. On image data, PCA yields global Fourier filters, due to the stationarity of image statistics (Field 1994). PCA has the nice property that it allows one to optimally (linearly) reduce the dimension by selecting only a subset of the components of $\boldsymbol{z} = \mathbf{V}_{\mathrm{PCA}}\,\boldsymbol{x}$. For grayscale patches, reducing the dimension this way while whitening is essentially equivalent to a combined low-pass and whitening filter which has been shown to be an optimal whitening filter assuming a finite noise level (Atick and Redlich 1992), and has been used for example in Olshausen and Field (1996, 1997). In addition, reducing the dimension also lowers the computational costs (running time and memory consumption) of the ICA estimation. We will exploit this property when working on colour and stereo data.

Having preprocessed the data, the goal of ICA is a transform $\mathbf{W}$ which minimizes the statistical dependences between the estimated sources

$$\hat{\boldsymbol{s}} = \mathbf{W}\boldsymbol{z} = \mathbf{W}\mathbf{V}_{\mathrm{PCA}}\boldsymbol{x} = \mathbf{W}\mathbf{D}_n^{-1/2}\mathbf{E}_n^{\mathrm{T}}\boldsymbol{x} \tag{5}$$

where we have by $\mathbf{D}_n$ denoted the diagonal matrix containing the n largest eigenvalues (of the correlation matrix $E\{\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$) and $\mathbf{E}_n$ the matrix with corresponding eigenvectors as columns. Optimally, the dependences are measured by the mutual information of the sources (Hyvärinen 1999a). However, to obtain a fast and simple algorithm, we constrain $\mathbf{W}$ to be orthogonal (i.e. our estimated sources are constrained to be uncorrelated) and approximate the mutual information as suggested in Hyvärinen (1999a). Thus, we use the FastICA algorithm (Hyvärinen and Oja 1997, Hyvärinen 1999a), with nonlinearity $g_1(u) = \tanh(u)$ (Hyvärinen 1999a) starting from a random orthogonal matrix $\mathbf{W}$. In essence, each iteration of the algorithm consists of updating each row $\boldsymbol{w}_i^{\mathrm{T}}$ of $\mathbf{W}$ by

$$\boldsymbol{w}_i := E\left\{\boldsymbol{z}g_1\left(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{z}\right)\right\} - E\left\{g_1'\left(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{z}\right)\right\}\boldsymbol{w}_i \tag{6}$$

followed by orthonormalization of the matrix through $\mathbf{W} := (\mathbf{W}\mathbf{W}^{\mathrm{T}})^{-1/2}\mathbf{W}$. After convergence, the estimated basis is constructed as

$$\mathbf{A} = \mathbf{E}_n\mathbf{D}_n^{1/2}\mathbf{W}^{\mathrm{T}} \tag{7}$$

and each column $\boldsymbol{a}_i$ of $\mathbf{A}$ can be identified as the basis patch of one independent component, as in figure 1.

2.2. The ICA basis: properties and relation to neural receptive fields

A typical ICA basis of image patches, when trained on natural grayscale images, is given in figure 2. The features (basis vectors) are Gabor-like filters at various positions, orientations, spatial frequencies and phases. One of the features codes the average gray level (dc-component).

Note also the resemblance to 'wavelets' (Mallat 1989, Daubechies 1992) (and related multiresolution representations) which have received significant attention lately as useful features in image processing. Indeed, one can argue that the reason wavelets are so successful

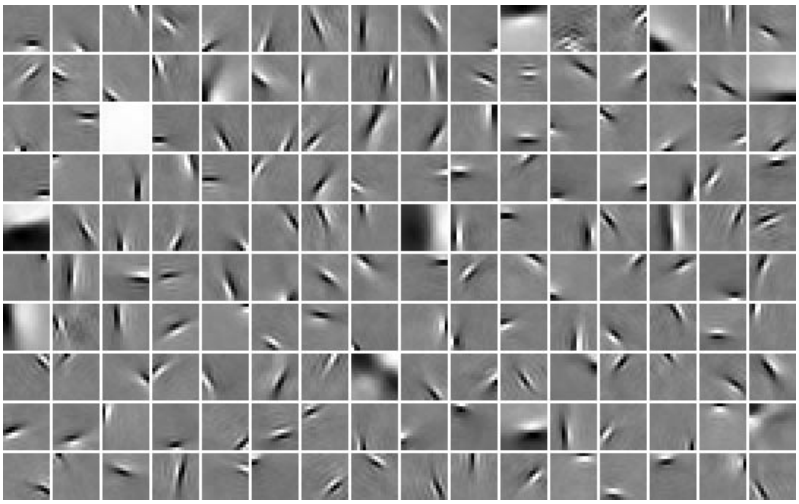

Figure 2. ICA basis of patches from grayscale images. 16-by-16 patches were sampled, and the dimension was reduced to 160 as described in section 2. The data was whitened and the FastICA algorithm was used to estimate the mixing matrix $\mathbf{A}$. Each patch corresponds to one column $\boldsymbol{a}_i$ of the estimated mixing matrix.

is precisely the fact that they form a sparse representation for the images (Mallat 1989, Hurri *et al* 1997).

The striking resemblance of the ICA features to receptive fields of neurons in the primary visual cortex suggests that the neurons do indeed perform some type of independent component analysis, and that the receptive fields are optimized for processing natural images (Olshausen and Field 1996, Bell and Sejnowski 1997). It is important to understand that this does not mean that the 'learning rule' or 'algorithm' actually operating in the cortex is anything like ours; rather it lets us understand the *purpose* of the computations as finding independent components of the input data. In the terminology of Marr (1982), we model the computational level instead of the algorithmic or implementation levels.

Note that if we want to establish a firm connection between ICA results and receptive fields we have to select our input images to be as close as possible to those which neurons would receive as input. Luckily, it seems that the ICA basis is not very sensitive to the particular set of natural images used, as earlier work on ICA for feature extraction of natural images has given qualitatively quite similar features using different data sets (Olshausen and Field 1997, Bell and Sejnowski 1997, Hurri *et al* 1997, van Hateren and van der Schaaf 1998). Thus, the selection of a reasonable dataset does not seem to be all that difficult.

An additional issue is whether one should compare the independent component filters or the basis vectors to the receptive fields of primary visual cortical neurons. Although previous studies (van Hateren and van der Schaaf 1998, van Hateren and Ruderman 1998) have compared the filters (rows of the separating matrix $\mathbf{WV}$) to measured receptive fields we feel that it is perhaps often more useful to look at the basis vectors (columns of $\mathbf{A}$). Although the filters can be equated with the feedforward connections of the neurons, the basis vectors form the 'optimal stimuli', in the sense that basis vector $\boldsymbol{a}_i$ gives a non-zero response only in unit i. This means that the *relative* response of that unit is maximized for input $\boldsymbol{a}_i$†. In addition, the visualization of the filters is not nearly as straightforward as visualization of the optimal stimuli, at least in the case of colour data.

A final question is how well the receptive fields of simple cells can be described using linear models. Simple cell responses are certainly not completely linear; they show significant nonlinearities such as rectification and response saturation. However, it seems that such nonlinearities can be thought of as operating 'on top' of a linear representation: models employing an underlying linear mechanism followed by some form of rectification and gain control have been quite successful (Heeger 1992, DeAngelis *et al* 1993b, Carandini *et al* 1997). Thus, it makes sense to compare linear ICA features to the representation given by simple cells.

3. Colour image experiments

In this section, we extend the grayscale ICA image model to include colours. Thus, for each pixel we have three values (red, green, blue) instead of one (grayscale). The corresponding ICA model is illustrated in figure 3. First, we discuss the selection of data, then we analyse its second-order statistics and finally we show the features found using ICA.

3.1. Choice of data

As discussed in section 2, we should select as input data images as 'natural' as possible if we wish to make any connection between our results and the properties of neurons in the visual

† This argument is even stronger if one takes into account normalization mechanisms (Heeger 1992, Carandini *et al* 1997) which suppress neuronal responses when a large number of neurons are simultaneously activated. Such mechanisms may favour inputs which activate only single (or few) units.

cortex. When analysing colours, the spectral composition of the images becomes important in addition to the spatial structure.

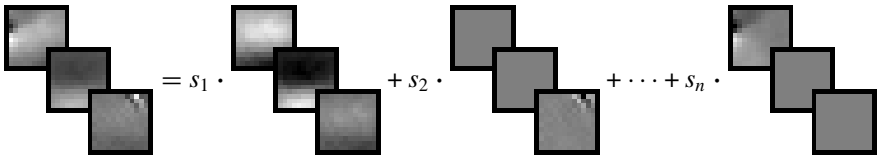

Figure 3. The colour image ICA model. Again, we model the data as a linear combination of 'basis patches', as in figure 1. Here, each patch consists of the three colour planes (red, green and blue), shown separately to clearly illustrate the linear model. In ICA, the purpose is to find the basis which minimizes the statistical dependences between the coefficients s_i.

It is clear that the colour content of images varies widely with the environment in which the images are taken. Thus, we do not pretend to find some universally optimal basis in which to code all natural colour images. Rather, we seek the general qualitative properties of an ICA decomposition of such images. In other words, we hope to find answers to questions such as the following. How are colours coded in such a basis; separate from, or multiplexed onto achromatic channels? What kind of spatial configuration do colour-coding basis vectors have?

Neurons of course receive their information ultimately from the outputs of the cones in the retina. Thus our data should consist of the hypothetical outputs of the three types of cones in response to our images. However, any three linear combinations of these outputs is just as good an input data, since we are applying ICA: linearly transforming the data transforms the mixing matrix, but does not alter the independent components (sources).

We choose to use standard red/green/blue (RGB) values as inputs, assuming the transformation to cone outputs to be roughly linear. This has the advantage that the features found are directly comparable to features currently in use in image processing operations, such as compression or denoising, and could straightforwardly be applied in such tasks. The drawback of using RGB values as inputs is of course that any nonlinearities inherent in the conversion from RGB to cone responses will affect the ICA result and a comparison to properties of neurons may not be warranted. To test the effect of nonlinearities, we have experimented with transforming the RGB values using the well-known gamma-nonlinearity† of CRTs. This did not qualitatively change the results, and therefore we are confident that our results would be similar if we had used estimated cone outputs as inputs.

Our main data consists of colour versions of natural scenes (depicting forest, wildlife, rocks, etc) which we have used in previous work as well (Hyvärinen 1999b, Hyvärinen and Hoyer 2000). The data is in the form of 20 RGB images (of size 384 × 256 pixels) in standard TIFF format; an example is given in figure 4.

3.2. Preprocessing

From the images, a total of 50 000 12 × 12 pixel image patches were sampled randomly. Since each channel yields 144 pixels, the dimensionality was now 3 × 144 = 432. Next, the mean value of each component was subtracted from that component, centring the dataset on the origin. As mentioned in section 2, this is a standard preprocessing step in ICA.

Then, we calculated the correlation matrix and its eigenvectors. These are shown in figure 5. The eigenvectors consist of global features, resembling 2D Fourier bases.

† The gamma-nonlinearity is the most significant nonlinearity of the CRT monitor. After gamma-correction the transform from RGB to cone responses is roughly linear (see the appendix in Wandell 1995).

Figure 4. One of the colour images used in the experiments.

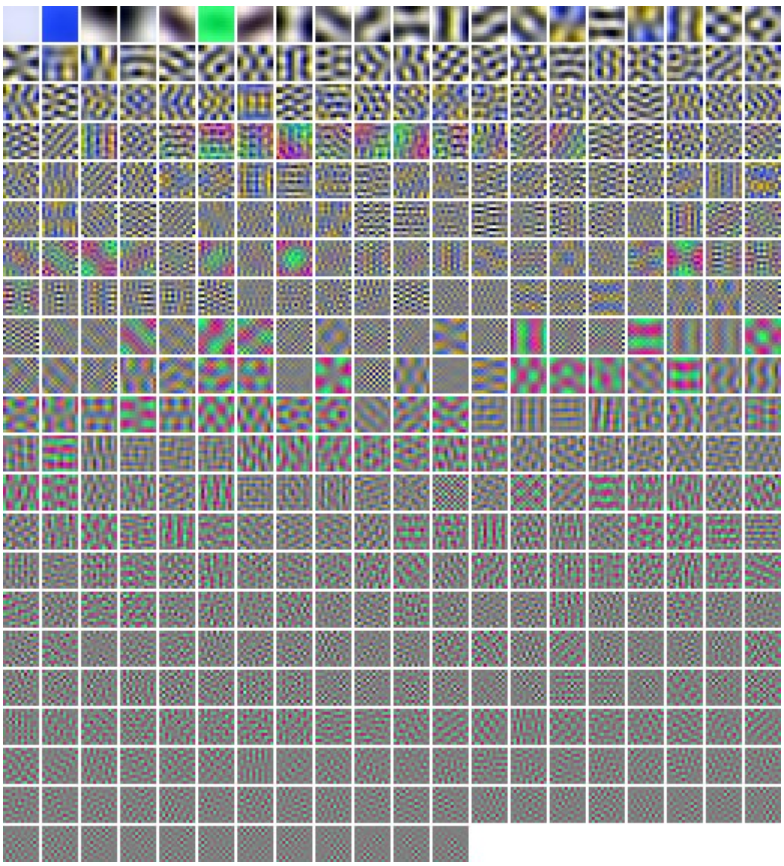

Figure 5. PCA basis of colour images. These are the eigenvectors of the correlation matrix of the data, from left-to-right and top-to-bottom in order of decreasing corresponding eigenvalues. As explained in the main text, we projected the data on the first 160 principal components (top eight rows) before performing ICA.

Figure 6. The colour hexagon used for analysing the colour content of the PCA and ICA basis patches. The hexagon is the projection of the RGB cube onto a plane orthogonal to the luminance ($R + G + B$) vector. Thus, achromatic RGB triplets map to the centre of the hexagon while highly saturated ones are projected close to the edges.

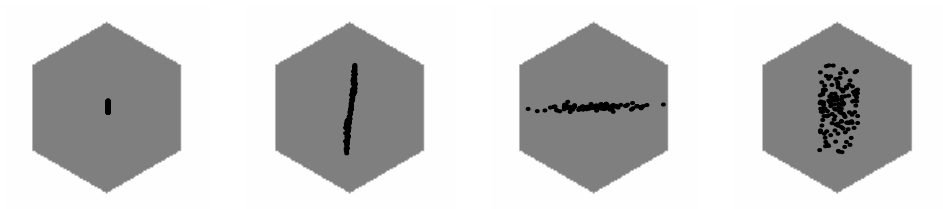

Figure 7. Colour content of four PCA filters. From left to right: component numbers 3, 15, 432, and 67. All pixels of each filter have been projected onto the colour hexagon shown in figure 6. See the main text for a discussion of the results.

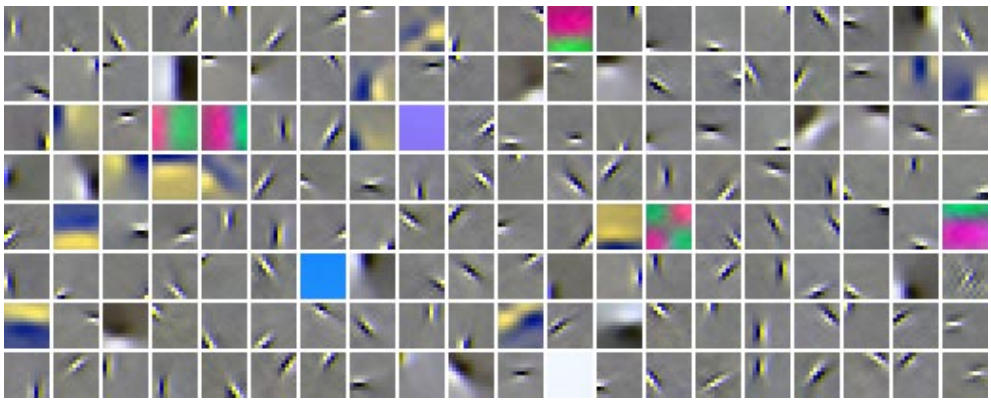

Figure 8. ICA basis of colour images. Each patch corresponds to one of the columns $\boldsymbol{a}_i$ of the estimated mixing matrix. Note that each patch is equally well represented by its negation, i.e. switching each pixel to its opponent colour in any one patch is equivalent to changing the sign of $\boldsymbol{a}_i$ and does not change the ICA model (assuming components with symmetric distributions).

The variance decreases with increasing spatial frequency, and when going from grayscale to blue/yellow to red/green features†. These results were established by Ruderman *et al* (1998), who used hyperspectral images as their original input data.

To analyse the colour content of the PCA filters in more detail, we will show the pixels of a few filters plotted in a coloured hexagon. In particular, each pixel (RGB-triplet) is projected onto a plane given by

$$R + G + B = \text{constant}. \tag{8}$$

† It should be noted that chromatic aberration in the eye might have an effect of additionally reducing signal energy at high spatial frequencies.

In other words, the luminance is ignored, and only the colour content is used in the display. Figure 6 shows the colours in this hexagon. Note that this is a very simple 2D projection of the RGB colour cube and should not directly be compared to any psychophysical colour representations.

Figure 7 shows a bright/dark filter (number 3 in figure 5), a blue/yellow filter (number 15), a red/green filter (number 432, the last one), and a mixture (number 67). Most filters are indeed exclusively opponent colours, as was found in Ruderman *et al* (1998). However, there are also some mixtures of these in the transition zones of the main opponent colours.

As described earlier, we project the data onto the n first principal components before whitening (we have experimented with $n = 100$, 160, 200, and 250). This is done for two reasons. First, something similar is probably done in real neurons, as amplifying directions with small variance would be disastrous in terms of signal-to-noise ratio (Atick and Redlich 1992). Second, the dimension is dropped to lower computational costs.

As can be seen from figure 5, dropping the dimension mostly discards the blue/yellow features of high spatial frequency and the red/green features of medium to high frequency. This could give a hint as to why the blue/yellow and the red/green systems have a much lower resolution than the bright/dark system, as has been observed in psychophysical experiments (Mullen 1985). As explained in section 2, the projected data is now whitened, and the FastICA algorithm run on the whitened data.

3.3. Results and discussion

The columns $\boldsymbol{a}_i$ of the estimated ICA mixing matrix $\mathbf{A}$, as given by (7) with $n = 160$, are shown in figure 8. Before analysing it further, one should consider the following question. How does the basis shown relate to the receptive fields we would have obtained had we used cone outputs as input data? A moment of reflection should convince you that since we have here used the RGB values to generate the image for your visual system, we are actually providing what would have been the optimal stimuli for our 'model neurons' had they learnt from cone outputs. The basis can thus directly be compared to the receptive fields of real neurons. (See section 2.2 for a discussion of whether to compare receptive fields to the basis vectors or the filters.)

Examining figure 8 closely reveals that the features found are very similar to earlier results (Olshausen and Field 1996, Bell and Sejnowski 1997) on grayscale image data, i.e. the patches resemble Gabor functions. Note that most units are (mainly) achromatic, so they only represent brightness (luminance) variations. This is in agreement with the finding that a large part of the neurons in the primary visual cortex seem to respond equally well to different coloured stimuli, i.e. they are not selective to colour (Hubel and Wiesel 1968, Livingstone and Hubel 1984). In addition, there is a small number of red/green and blue/yellow patches. These are also oriented, but of much lower spatial frequency, similar to the grayscale patches of lowest frequency. One could think that the low frequency patches together form a 'colour' (including brightness) system, and the high-frequency grayscale patches a channel analysing form. Also note that the average colour (dc-value) of the patches is represented by three separate basis vectors, just as the average brightness in an ICA decomposition of grayscale images is usually separate from the other basis vectors.

We now show typical ICA basis patches plotted in the colour hexagon (figure 9), as we did with the PCA basis. The figure shows a bright/dark patch, a blue/yellow patch, and a red/green patch. There were no 'mixtures' of the type seen for PCA; in other words each patch clearly belonged to one of these groups. (Note that the bright/dark patches also contained blue/yellow to a quite small degree.)

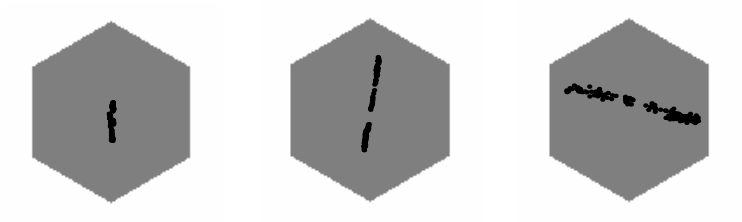

Figure 9. Colour content of three ICA filters, projected onto the colour hexagon of figure 6. From left to right: numbers 24, 82, and 12.

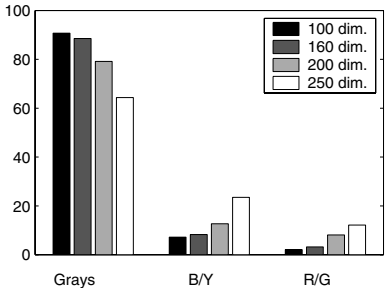

Figure 10. Percentages of achromatic, blue/yellow and red/green basis vectors for different numbers of retained PCA components (100, 160, 200 and 250). (In each case, the three patches giving the mean colour have been left out of this count.)

The dominance of bright/dark patches is largely due to the dimension reduction performed while whitening. To test the dependence of the group sizes on the value of n used, we estimated the ICA basis for different values of n and counted the group sizes in each case. The results can be seen in figure 10. Clearly, when n is increased, the proportion of colour-selective units increases. However, even for the case of keeping over half of the dimensions of the original space ($n = 250$), the bright/dark features still make up over 60% of all units.

Another thing to note is that each ICA basis patch is 'double-opponent': for blue/yellow patches, stimulating with a blue spot always gives an opposite sign in the response compared to stimulating with a yellow spot. Red/green and bright/dark features behave similarly. This is in fact a direct consequence of the linear ICA model. It would be impossible to have completely linear filters function in any other way.

Although early results (Livingstone and Hubel 1984) on the chromatic properties of neurons suggested that most colour-sensitive cells were unoriented, and exhibited centre-surround receptive fields, more recent studies have indicated that there are also oriented colour-selective neurons (Ts'o and Gilbert 1988). The fact that our colour features are mostly oriented is thus at least in partial agreement with neurophysiological data.

In any case, there is some agreement that most neurons are not selective to chromatic contrast, but rather they are more concerned about form (Hubel and Wiesel 1968, Livingstone and Hubel 1984, Ts'o and Roe 1995). Our basis is in agreement with these findings. In addition, the cytochrome oxidase blobs which have been linked to colour processing (Livingstone and Hubel 1984) have also been associated with low spatial frequency tuning (Tootell *et al* 1988, Shoham *et al* 1997). In other words, colour selective cells should be expected to be tuned to lower spatial frequencies. This is also seen in our features.

As stated earlier, we do not pretend that our main image set is representative of all natural environments. To check that the results obtained do not vary wildly with the image set used, we have performed the same experiments on another dataset: single-eye colour versions of the 11 stereo images described in section 4.1. The found ICA basis (not shown) is in most aspects quite similar to that shown in figure 8: features are divided into bright/dark, blue/yellow and red/green channels, of which the bright/dark group is the largest, containing Gabor-like filters of mostly higher frequency than the features coding colours. The main differences are that (a) there is a slightly higher proportion of colour-coding units, and (b) the opponent colours they code are slightly shifted in colour space from those found from our main data. In other words, the qualitative aspects, answering questions such as those proposed in section 3.1, are quite similar. However, quantitative differences do exist.

4. Stereo image experiments

Another interesting extension of the basic grayscale image ICA model can be made by modelling stereopsis, the extraction of depth cues from binocular disparity. Now, our artificial neurons are attempting to learn the dependences of corresponding patches from natural stereo images. The model is shown in figure 11.

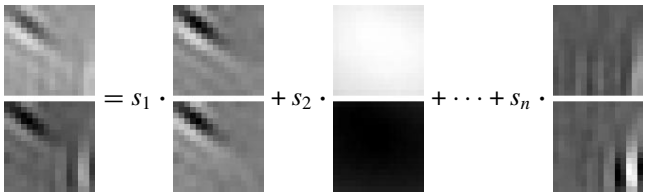

Figure 11. The ICA model for corresponding stereo image patches. The top row contains the patches from the left image and the bottom row corresponding patches from the right image. Just as for gray and colour patches, we model the data as a linear combination of basis vectors with independent coefficients.

4.1. Choice of data

Again, the choice of data is an important step for us to get realistic results. Previous studies have used different approaches. In some early work, a binocular correlation function was estimated from actual stereo image data, and subsequently analysed (Li and Atick 1994). In addition, at least one investigation of receptive field development used artificially generated disparity from monocular images (Shouval *et al* 1996). We have chosen to use 11 images from a commercial collection† of stereo images of natural scenes; a typical image is given in figure 12.

To simulate the workings of the eyes, we selected five focus points at random from each image and estimated the disparities at these points. We then randomly sampled 16×16 pixel corresponding image patches in an area of 300×300 pixels centred on each focus point, obtaining a total of 50 000 samples. Because of the local fluctuations in disparity (due to the 3D imaging geometry) corresponding image patches often contained similar, but horizontally shifted features; this is of course the basis of stereopsis.

Note that in reality the 'sampling' is quite different. Each neuron sees a certain area of the visual field which is relatively constant with respect to the focus point. Thus a more realistic sampling would be to randomly select 50 000 focus points and from each take corresponding image patches at some given constant positional offset. However, the binocular matching is computationally slow and we thus opted for the easier approach, which should give the same distribution of disparities.

† Available at `http://members.home.net/holographics/cd~1.htm`.

Figure 12. One of the stereo images used in the experiments. The left image should be seen with the left eye, and the right image with the right eye (i.e. uncrossed viewing).

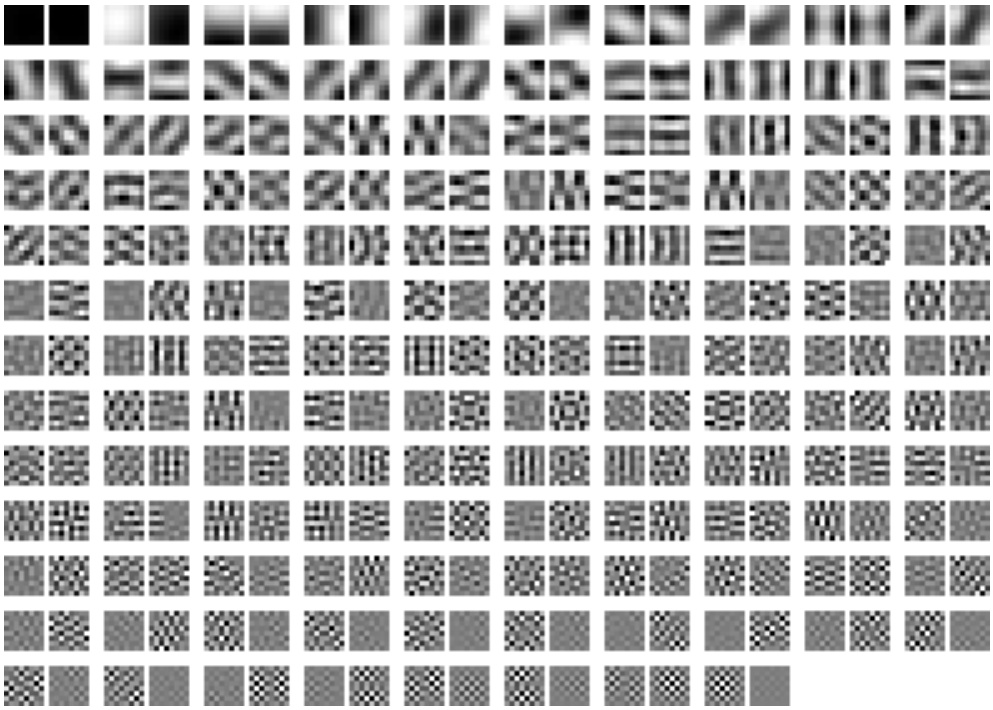

Figure 13. PCA basis of stereo images, i.e. the eigenvectors of the correlation matrix of the data, from left to right and top to bottom in order of decreasing corresponding eigenvalues. See the main text for a discussion.

4.2. Preprocessing

The same kind of preprocessing was used in these experiments as for colour, in section 3. Since each sample consisted of corresponding left and right 16×16 patches our original data was 512-dimensional. First, the mean was removed from each component, to centre the data on the origin. Next, we calculated the correlation matrix of the data, and its eigenvalue decomposition. Due to space limitations we show here (in figure 13) the principal components for a window size of 8×8 pixels (the result for 16×16 is qualitatively very similar).

The most significant feature is that the principal components are roughly ordered according to spatial frequency, just as in PCA on standard (monocular) image patches. However, in addition, early components (low spatial frequency) are more binocular than late ones (high frequency). Also note that binocular components generally consist of patches of identical or opposite phases. This is in agreement with the binocular correlation function described in Li and Atick (1994).

As before, we select the first 160 principal components for further analysis by ICA. Again, this is plausible as a coding strategy for neurons, but it is mainly done to lower the computational expenses and thus the running time and memory consumption. Due to the structure of the correlation matrix, dropping the dimension to 160 is similar to low-pass filtering.

4.3. Results and discussion

Figure 14 shows the estimated basis vectors after convergence of the FastICA algorithm. Each pair of patches represents one basis vector $\boldsymbol{a}_i$. First, note that the pairs have varying degrees of binocularity. Many of our 'model neurons' respond equally well to stimulation from both eyes, but there are also many which respond much better to stimulation of one eye than to stimulation of the other. This is shown quantitatively in figure 15, which gives an 'ocular-dominance' histogram of the features.

The histogram depends strongly on the area of the sampling around the focus points (which in these experiments was 300×300 pixels). Sampling a smaller area implies that the correlation between the patches is higher and a larger number of features fall into the middle bin of the histogram. In theory, if we chose to sample only exactly at the fixation point, we would obtain (ignoring factors such as occlusion) identical left–right image patches; this would in turn make all basis vectors completely binocular with identical left–right patches, as there would be no signal variance in the other directions of the data space. On the other hand, sampling a larger area leads to a spreading of the histogram towards the edge bins. As the area gets larger, the dependences between the left and right patches get weaker. In the limit of unrelated left and right windows, all features fall into bins 1 and 7 of the histogram. This was confirmed in the experiments (results not shown).

Taking a closer look at the binocular pairs reveals that for most pairs the left patch is similar to the right patch both in orientation and spatial frequency. The positions of the features inside the patches are close, when not identical. In some pairs the phases are very similar, while in others they are quite different, even completely opposite. These properties make the features sensitive to different degrees of binocular disparity. Identical left–right receptive fields make the features most responsive to zero disparity, while receptive fields that are identical, except for a phase reversal, show strong inhibition (a response smaller than the 'base-line' response given by an optimal monocular stimulus) to zero disparity.

To analyse the disparity tuning we first estimated several ICA bases using different random number seeds. We then selected only relatively high frequency, well-localized, binocular basis vectors which had a clear Gabor filter structure. This was necessary because filters of low

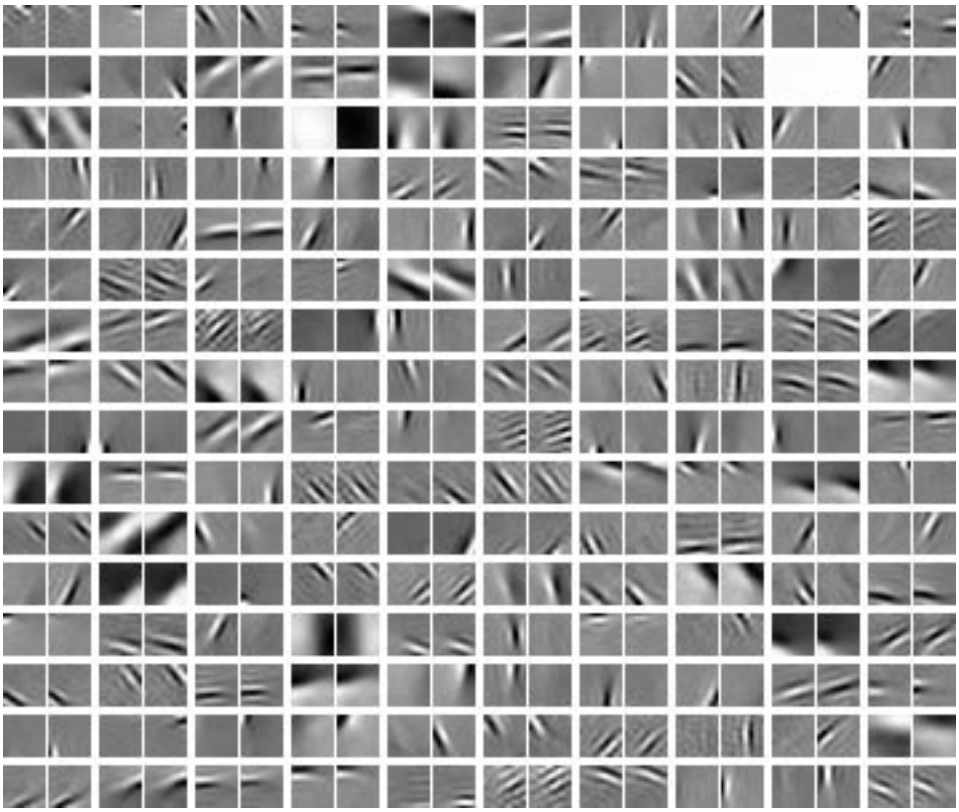

Figure 14. ICA basis of stereo images. Each pair of patches represents one basis vector $\boldsymbol{a}_i$ of the estimated mixing matrix $\mathbf{A}$. Note the similarity of these features to those obtained from standard image data. In addition, these exhibit various degrees of binocularity and varying relative positions and phases.

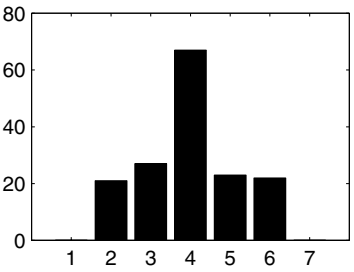

Figure 15. Ocular dominance histogram of the ICA features. For each pair, we calculated the value of $(\|\boldsymbol{a}_{\text{left}}\| - \|\boldsymbol{a}_{\text{right}}\|)/(\|\boldsymbol{a}_{\text{left}}\| + \|\boldsymbol{a}_{\text{right}}\|)$, and used the bin boundaries $[-0.85, -0.5, -0.15, 0.15, 0.5, 0.85]$ as suggested in Shouval *et al* (1996). Although many units were quite monocular (as can be seen from figure 14), no units fell into bins 1 or 7. This histogram is quite dependent on the sampling window around the fixation points, as discussed in the main text.

spatial frequency were not usually well confined within the patch and thus cannot be analysed as complete neural receptive fields. The set of selected basis vectors is shown in figure 16.

For each stereo pair, we presented an identical stimulus at different disparities to both the left and right parts of the filter corresponding to the pair. For each disparity, the maximum over translations was taken as the response of the pair at that disparity. This gave a disparity tuning curve. For stimuli we used the optimal stimuli (basis vectors) themselves, first presenting the left patch of the pair to both 'eyes', then the right. The tuning curves were usually remarkably similar, and we took the mean of these as the final curve.

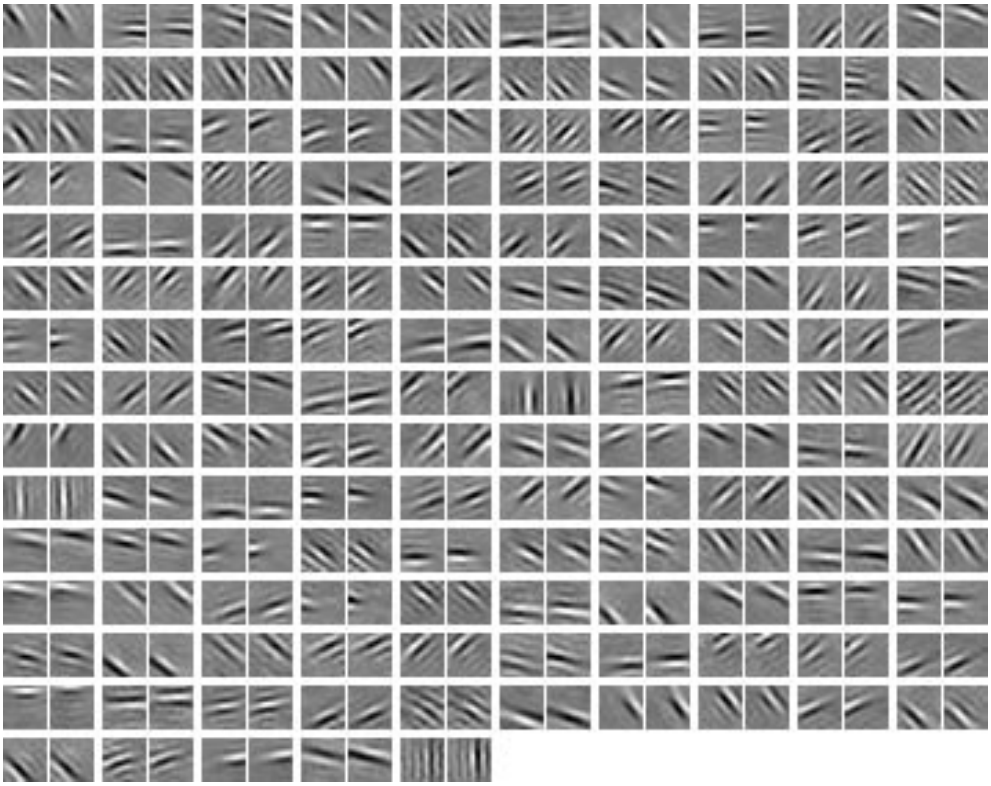

Figure 16. Units selected for disparity tuning analysis. These were selected from bases such as the one in figure 14 on the basis of binocularity, frequency content and localization (only well-localized Gabor filters were suitable for further analysis).

We then classified each curve as belonging to one of the types, 'tuned excitatory', 'tuned inhibitory', 'near' or 'far', which have been identified in physiological experiments (Poggio and Fischer 1977, Fischer and Kruger 1979, LeVay and Voight 1988). Tuned excitatory units showed a strong peak at zero, usually with a smaller inhibition at either side. Tuned inhibitory units on the other hand showed a marked inhibition (cancelling) at zero disparity, with excitation at small positive or negative disparities. Features classified as 'near' showed a clear positive peak at crossed (positive) disparity while those grouped as 'far' showed a peak for uncrossed (negative) disparity. Some tuning curves that did not clearly fit any of these classes were grouped into 'others'.

In figure 17 we give one example from each class. The basis vectors and the corresponding tuning curves are shown. It is fairly easy to see how the organization of the patches gives the tuning curves. The tuned excitatory (top) unit has almost identical left–right profiles and thus shows a strong preference for stimuli at zero disparities. The tuned inhibitory (second) unit has nearly opposite polarity patches which implies strong inhibition at zero disparity. The right receptive field of the near (third) unit is slightly shifted (positional offset) to the left compared with the left field, giving it a positive preferred disparity. On the other hand, the far unit (bottom) has an opposite positional offset and thus responds best to negative disparities.

Figure 18 shows the relative number of units in the different classes. Note that the most common classes are 'tuned excitatory' and 'near'. One would perhaps have expected a greater

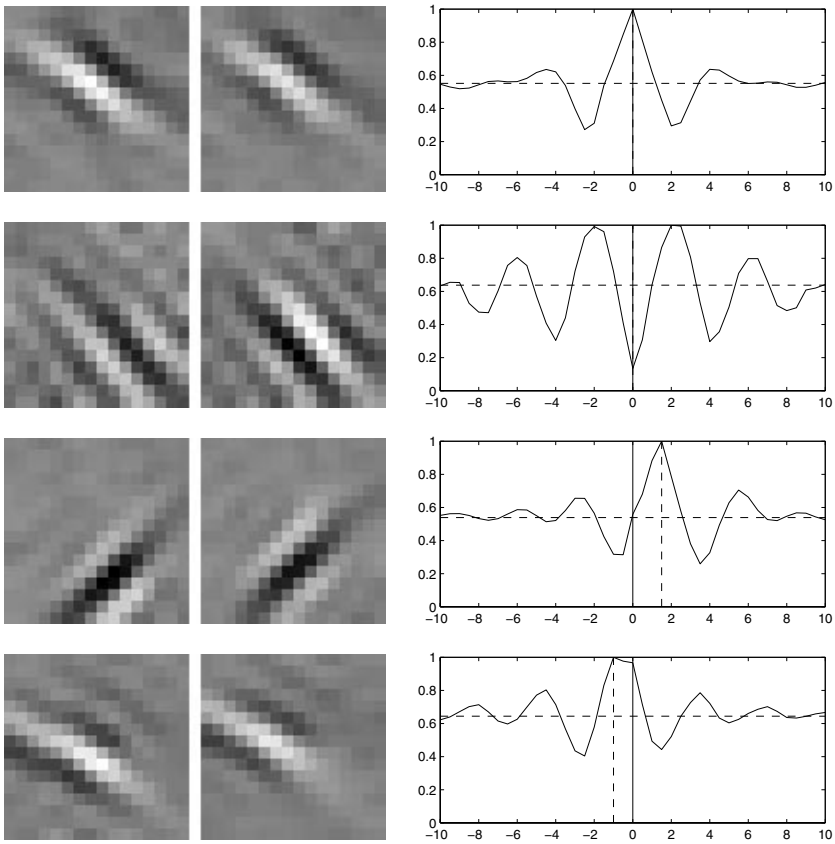

Figure 17. Disparity tuning curves for units belonging to different classes. Top row: a 'tuned excitatory' unit (number 4 in figure 16). Second row: a 'tuned inhibitory' unit (12). Third row: a 'near' unit (38). Bottom row: a 'far' unit (47). Crossed disparity ('near') is labelled positive, and uncrossed ('far') negative in the figures. The horizontal dotted line gives the 'base-line' response (the optimal response to one eye only) and the vertical dotted line the position of maximum deviation from that response.

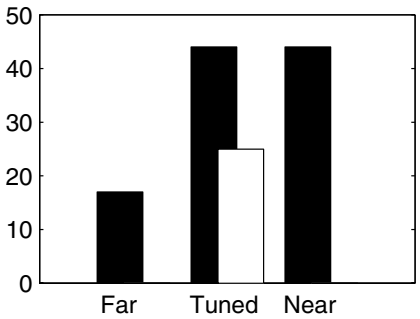

Figure 18. Disparity tuning histogram. The histogram shows the relative amounts of 'tuned excitatory' (44), 'near' (44), 'far' (17) units (in black) and 'tuned inhibitory' units (25) in white. Not shown are those which did not clearly fit into any of these categories (15).

dominance of the tuned excitatory over the other groups. The relative number of tuned versus untuned units probably depends to a great deal on the performance of the disparity estimation algorithm in the sampling procedure. We suspect that with a more sophisticated algorithm (we have used a very simple window-matching technique) one would get a larger number of tuned cells. The clear asymmetry between the 'near' and 'far' groups is probably due to the much larger range of possible disparities for near than for far stimuli: disparities for objects closer than fixation can in principle grow arbitrarily large whereas disparities for far objects are limited (Barlow *et al* 1967).

It is important to note that completely linear units (simple cells) cannot have very selective disparity tuning. Also, since the disparity tuning curves vary with the stimulus, the concept 'disparity tuning curve' is not even well-defined (Zhu and Qian 1996). However, disparity tuning is still measurable so long as one keeps in mind that the curve depends on the stimulus. Our tuning curves are 'simulations' of experiments where a moving stimulus is swept across the receptive field at different binocular disparities, and the responses of the neuron in question are measured. As such, it is appropriate to use the optimal stimuli (basis vectors) as input. To obtain stimulus-invariant disparity tuning curves (as well as more complex binocular interactions than those seen here) one would need to model nonlinear (complex) cells.

Overall, the properties of the found features correspond quite well to those of receptive fields measured for neurons in the visual cortex. The features show varying degrees of ocular dominance, just as neuronal receptive fields (Hubel and Wiesel 1962). Binocular units have interocularly matched orientations and spatial frequencies, as has been observed for real binocular neurons (Skottun and Freeman 1984). It is easy by visual inspection to see that there exist both interocular position and phase differences, which seems to be the case for receptive fields of cortical neurons (Anzai *et al* 1999). Finally, simulated disparity tuning curves of the found features are also similar to tuning curves measured in physiological experiments (Poggio and Fischer 1977).

5. Relation to previous work

Numerous models have been proposed for the development of visual receptive fields, and the neuronal learning of input statistics. Early work concentrated on the second-order statistics of the input, giving solutions closely related to PCA (Oja 1982, Miller 1990, Hancock *et al* 1992). In addition, many authors have developed optimal information processing frameworks assuming Gaussian input data and additive Gaussian noise (Linsker 1988, Atick and Redlich 1990, van Hateren 1992).

Recently, however, it has been argued that natural image data is actually far from Gaussian (Field 1994). Thus, the second-order statistics fail to describe important aspects of the data, and one must use higher-order information to get an efficient representation. ICA is a fundamental method for such learning, and the match of the ICA representation to neuronal receptive fields is impressive, if not complete (van Hateren and van der Schaaf 1998).

Although there have been numerous studies of learning grayscale receptive fields, not many have attempted the analysis of colour or stereopsis. There has been some work concerning the second-order statistics of colour (Atick *et al* 1992, van Hateren 1993, Ruderman *et al* 1998). In addition, coloured input was used in Barrow *et al* (1996) to emerge a topographic map of receptive fields. Again, that work basically concerns only the second-order structure of the data, as the correlation-based learning used relies only on this information. The current work is thus the first (to the knowledge of the authors) to work with higher-order statistics of colour images†.

† The authors have recently learned of somewhat similar, concurrent work reported in Lee *et al* (2000).

Emerging receptive fields from stereo input have been considered in Li and Atick (1994), Shouval *et al* (1996), and Erwin and Miller (1996, 1998). As with colour, most studies have explicitly or implicitly used only second-order statistics (Li and Atick 1994, Erwin and Miller 1996, 1998). The exception is Shouval *et al* (1996) which used the BCM learning rule (Bienenstock *et al* 1982) which is a type of projection pursuit learning closely linked to ICA. The main difference between their work and ours is that we use data from actual stereo images whereas they used horizontally shifted (misaligned) data from regular images. In addition, we estimate a complete basis for the data, whereas they studied only single receptive fields.

6. Conclusions

We have investigated the use of independent component analysis for decomposing natural colour and stereo images. ICA applied to colour images yields basis vectors which resemble Gabor functions, with most features achromatic, and the rest red/green or blue/yellow opponent. When ICA is applied on stereo images we obtain feature pairs which exhibit various degrees of ocular dominance and are tuned to various disparities.

These results are significant for two reasons. First, the features learned by ICA could be straightforwardly applied in denoising, compression, or pattern recognition of colour or stereo data. In each of these tasks it is important to model the statistical structure of the data; ICA has been successfully used to model that structure (Hyvärinen 1999b). Second, ICA can be used to model computational properties of V1 cells. The similarity of the ICA features to optimal stimuli measured for neurons in the primary visual cortex using single-cell recordings suggests that these neurons perform some form of redundancy reduction, as proposed by Barlow (1989). It seems likely that information processing strategies successful in the primary visual cortex would also be useful in higher visual processing, and indeed in the processing of other sensory signals; thus it seems probable that ICA or related methods could be applied in modelling these functions as well.

Acknowledgments

The authors would like to thank Jukka Häkkinen, Jari Laarni, Pentti Laurinen, Bruno Olshausen, Tarja Peromaa and Harri Valpola for stimulating and useful discussions. In addition, we wish to thank the anonymous reviewers for suggestions which helped to improve the clarity of this paper.

References

Amari S, Cichocki A and Yang H 1996 A new learning algorithm for blind source separation *Advances in Neural Information Processing Systems 8* (Cambridge, MA: MIT Press) pp 757–63

Anzai A, Ohzawa I and Freeman R D 1999 Neural mechanisms for encoding binocular disparity: Receptive field position vs. phase *J. Neurophysiol.* **82** 874–90

Atick J J, Li Z and Redlich A N 1992 Understanding retinal color coding from first principles *Neural Comput.* **4** 559–72

Atick J J and Redlich A N 1990 Towards a theory of early visual processing *Neural Comput.* **2** 308–20

——1992 What does the retina know about natural scenes? *Neural Comput.* **4** 196–210

——1993 Convergent algorithm for sensory receptive field development *Neural Comput.* **5** 45–60

Barlow H B 1989 Unsupervised learning *Neural Comput.* **1** 295–311

Barlow H B, Blakemore C and Pettigrew J D 1967 The neural mechanism of binocular depth discrimination *J. Physiol.* **193** 327–42

Barrow H G, Bray A J and Budd J M L 1996 A self-organized model of 'color blob' formation *Neural Comput.* **8** 1427–48

Bell A and Sejnowski T 1995 An information-maximization approach to blind separation and blind deconvolution *Neural Comput.* **7** 1129–59
——1997 The 'independent components' of natural scenes are edge filters *Vis. Res.* **37** 3327–38
Bienenstock E L, Cooper L N and Munro P W 1982 Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex *J. Neurosci.* **2** 32–48
Carandini M, Heeger D J and Movshon J A 1997 Linearity and normalization in simple cells of the macaque primary visual cortex *J. Neurosci.* **17** 8621–44
Cardoso J-F and Laheld B H 1996 Equivariant adaptive source separation *IEEE Trans. Signal Process.* **44** 3017–30
Cichocki A and Unbehauen R 1996 Robust neural networks with on-line learning for blind identification and blind separation of sources *IEEE Trans. Circuits Syst.* **43** 894–906
Comon P 1994 Independent component analysis—a new concept? *Signal Process.* **36** 287–314
Dan Y, Atick J J and Reid R C 1996 Efficient coding of natural scences in the lateral geniculate nucleus: experimental test of a computational theory *J. Neurosci.* **16** 3351–62
Daubechies I 1992 *Ten Lectures on Wavelets* (Philadelphia, PA: Society for Industrial and Applied Math.)
DeAngelis G C, Ohzawa I and Freeman R D 1993a Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. I. General characteristics and postnatal development *J. Neurophysiol.* **69** 1091–117
——1993b Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. II. Linearity of temporal and spatial summation *J. Neurophysiol.* **69** 1118–35
DeValois R L, Lund E W and Hepler N 1982 The orientation and direction selectivity of cells in macaque visual cortex *Vis. Res.* **22** 531–44
Erwin E and Miller K D 1996 Modeling joint development of ocular dominance and orientation maps in primary visual cortex *Computational Neuroscience: Trends in Research 1995* ed J M Bower (New York: Academic) pp 179–84
——1998 Correlation-based development of ocularly-matched orientation and ocular dominance maps: determination of required input activities *J. Neurosci.* **18** 5908–27
Field D J 1994 What is the goal of sensory coding? *Neural Comput.* **6** 559–601
Fischer B and Kruger J 1979 Disparity tuning and binocularity of single neurons in cat visual cortex *Exp. Brain Res.* **35** 1–8
Hancock P J B, Baddeley R J and Smith L S 1992 The principal components of natural images *Network* **3** 61–72
Heeger D 1992 Normalization of cell responses in cat striate cortex *Vis. Neurosci.* **9** 181–98
Hubel D and Wiesel T 1962 Receptive fields, binocular interaction and functional architecture in the cat's visual cortex *J. Physiol.* **160** 106–54
——1968 Receptive fields and functional architecture of monkey striate cortex *J. Physiol.* **195** 215–43
Hurri J, Hyvärinen A and Oja E 1997 Wavelets and natural image statistics *Proc. Scandinavian Conf. on Image Analysis '97 (Lappenranta, Finland)*
Hyvärinen A 1999a Fast and robust fixed-point algorithms for independent component analysis *IEEE Trans. Neural Netw.* **10** 626–34
——1999b Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation *Neural Comput.* **11** 1739–68
——1999c Survey on independent component analysis *Neural Comput. Surv.* **2** 94–128
Hyvärinen A and Hoyer P O 2000 Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces *Neural Comput.* **12** at press
Hyvärinen A, Hoyer P O and Oja E 2000 Image denoising by sparse code shrinkage *Intelligent Signal Processing* ed S Haykin and B Kosko (Piscataway, NJ: IEEE) at press
Hyvärinen A and Oja E 1997 A fast fixed-point algorithm for independent component analysis *Neural Comput.* **9** 1483–92
Jutten C and Herault J 1991 Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture *Signal Process.* **24** 1–10
Karhunen J, Oja E, Wang L, Vigário R and Joutsensalo J 1997 A class of neural networks for independent component analysis *IEEE Trans. Neural Netw.* **8** 486–504
Lee T-W, Wachtler T and Sejnowski T J 2000 The spectral independent components of natural scenes *Proc. IEEE Int. Workshop on Biologically Motivated Computer Vision (BMCV'00) (Seoul, Korea)* pp 527–34
LeVay S and Voigt T 1988 Ocular dominance and disparity coding in cat visual cortex *Vis. Neurosci.* **1** 395–414
Li Z and Atick J J 1994 Efficient stereo coding in the multiscale representation *Network* **5** 157–74
Linsker R 1988 Self-organization in a perceptual network *Computer* **21** 105–17
Livingstone M S and Hubel D H 1984 Anatomy and physiology of a color system in the primate visual cortex *J. Neurosci.* **4** 309–56
Mallat S G 1989 A theory for multiresolution signal decomposition: The wavelet representation *IEEE Trans. Pattern*

Anal. Mach. Intell. **11** 674–93
Marr D 1982 *Vision* (New York: Freeman)
Miller K D 1990 Correlation-based models of neural development *Neuroscience and Connectionist Theory* ed M Gluck and D Rumelhart (Hillsdale, NJ: Lawrence Erlbaum) pp 267–353
Mullen K T 1985 The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings *J. Physiol.* **359** 381–400
Oja E 1982 A simplified neuron model as a principal component analyzer *J. Math. Biol.* **15** 267–73
——1997 The nonlinear PCA learning rule in independent component analysis *Neurocomputing* **17** 25–46
Olshausen B A and Field D J 1996 Emergence of simple-cell receptive field properties by learning a sparse code for natural images *Nature* **381** 607–9
——1997 Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vis. Res.* **37** 3311–25
Pajunen P 1998 Blind source separation using algorithmic information theory *Neurocomputing* **22** 35–48
Poggio G F and Fischer B 1977 Binocular interaction and depth sensitivity in striate and prestriate cortex of behaving rhesus monkey *J. Neurophysiol.* **40** 1392–1405
Ruderman D L, Cronin T W and Chiao C 1998 Statistics of cone responses to natural images: Implications for visual coding *J. Opt. Soc. Am.* A **15** 2036–45
Shoham D, Hübener M, Schulze S, Grinvald A and Bonhoeffer T 1997 Spatio-temporal frequency domains and their relation to cytochrome oxidase staining in cat visual cortex *Nature* **385** 529–33
Shouval H, Intrator N, Law C C and Cooper L N 1996 Effect of binocular cortical misalignment on ocular dominance and orientation selectivity *Neural Comput.* **8** 1021–40
Skottun B C and Freeman R D 1984 Stimulus specificity of binocular cells in the cat's visual cortex: Ocular dominance and the matching of left and right eyes *Exp. Brain Res.* **56** 206–16
Tootell R B H, Silverman M S, Hamilton S L, Switkes E and Valois R L D 1988 Functional anatomy of macaque striate cortex. V. Spatial frequency *J. Neurosci.* **8** 1610–24
Ts'o D Y and Gilbert C D 1988 The organization of chromatic and spatial interactions in the primate striate cortex *J. Neurosci.* **8** 1712–27
Ts'o D Y and Roe A W 1995 Functional compartments in visual cortex: Segregation and interaction *The Cognitive Neurosciences* ed M S Gazzaniga (Cambridge, MA: MIT Press) pp 325–37
van Hateren J H 1992 A theory of maximizing sensory information *Biol. Cybern.* **68** 23–29
——1993 Spatial, temporal and spectral pre-processing for colour vision *Proc. R. Soc.* B **251** 61–8
van Hateren J H and Ruderman D L 1998 Independent component analysis of natural image sequences yields spatiotemporal filters similar to simple cells in primary visual cortex. *Proc. R. Soc.* B **265** 2315–20
van Hateren J H and van der Schaaf A 1998 Independent component filters of natural images compared with simple cells in primary visual cortex *Proc. R. Soc.* B **265** 359–66
Wandell B A 1995 *Foundations of Vision* (Sunderland, MA: Sinauer Associates)
Zhu Y and Qian N 1996 Binocular receptive field models, disparity tuning, and characteristic disparity *Neural Comput.* **8** 1611–41