$$\Delta\mathbf{V} \propto (\mathbf{I}+\mathbf{V})(\mathbf{I}+\hat{\mathbf{y}}\mathbf{u}^{\mathrm{T}}). \qquad (16)$$

In terms of an individual feedback weight, v_{ij}, this rule is:

$$\Delta v_{ij} \propto \delta_{ij} + v_{ij} + u_j\left(\hat{y}_i + \sum_k v_{ik}\hat{y}_k\right) \qquad (17)$$

where $\delta_{ij} = 1$ when $i = j$, 0 otherwise. Thus, the feedback rule is also non-local, this time involving a backwards pass through the (recurrent) weights, of quantities, $\hat{y}_k$, calculated from the nonlinear output vector, **y**. Such a recurrent ICA system has been further developed for recovering sources which have been linearly convolved with temporal filters by Torkkola (1996) and Lee *et al.* (1997).

The non-locality of the algorithm is interesting when we come to consider the biological significance of the learned filters later in this paper.

METHODS

We took four natural scenes involving trees, leaves and so on* and converted them to greyscale byte values between 0 and 255. A training set, $\{\mathbf{x}\}$, was then generated of 17 595, 12 × 12 samples from the images. The training set was "sphered" by subtracting the mean and multiplying by twice the local symmetrical (zero-phase) whitening filter of equation (8):

$$\mathbf{x} \leftarrow 2\mathbf{W}_Z(\{\mathbf{x}\} - \langle\mathbf{x}\rangle) \qquad (18)$$

This removes both first- and second-order statistics from the data, and makes the covariance matrix of **x** equal to 4**I**. This is an appropriately scaled starting point for further training since infomax [equation (13)] on raw data, with the logistic function, $y_i = (1 + \exp(-u_i)^{-1}$, produces a **u**-vector which approximately satisfies $\langle\mathbf{u}\mathbf{u}^{\mathrm{T}}\rangle = 4\mathbf{I}$. Therefore, by prewhitening **x** in this way, we can ensure that the subsequent transformation, $\mathbf{u} = \mathbf{W}\mathbf{x}$, to be learnt should approximate an orthonormal matrix (rotation without scaling), roughly satisfying the relation $\mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}$ (Karhunen *et al.*, 1996). This **W** moves the solution along the decorrelating manifold from ZCA to ICA (see Fig. 2).

The matrix, **W**, is then initialized to the identity matrix, and trained using the logistic function version of equation (13), in which equation (12) evaluates as: $y_i = 1 - 2y_i$. The training was conducted as follows: 30 sweeps through the data were performed, at the end of each of which the order of the data vectors was permuted to avoid cyclical behaviour in the learning. During each sweep, the weights were updated only after every 50 presentations in order that the vectorized MATLAB code could be more efficient. The learning rate [proportionality constant in equation (13)] was set as follows: 21 sweeps at 0.001, and three sweeps at each of 0.0005, 0.0002 and 0.0001. This process took 2 hours running MATLAB on a Sparc-20 machine, though a reasonable result for 12 × 12 filters can be achieved in 30 min. To verify that the result was not affected by the starting condition of $\mathbf{W} = \mathbf{I}$, the training was repeated with several randomly initialized weight matrices, and also on data that were not prewhitened. The results were qualitatively similar, though convergence was much slower.

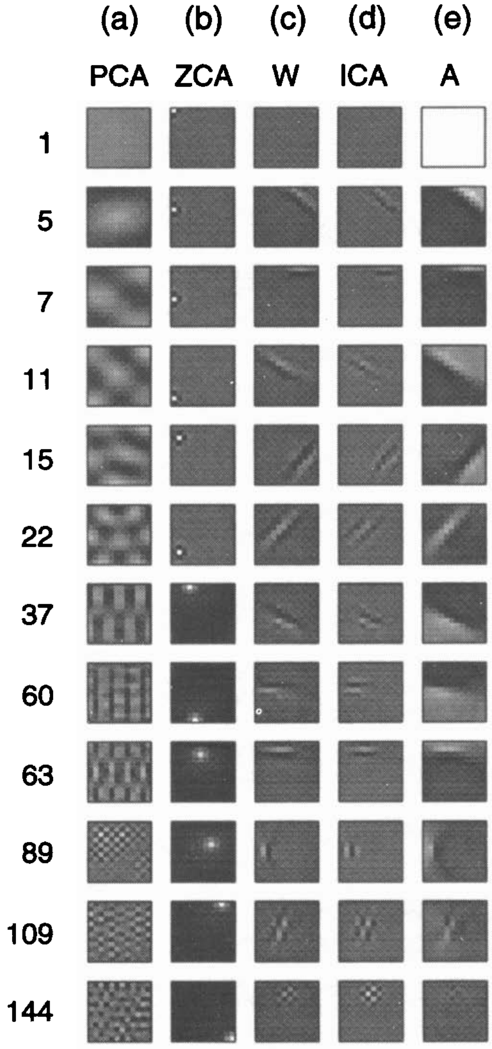

FIGURE 3. Selected decorrelating filters and their basis functions extracted from the natural scene data. Each type of decorrelating filter yielded 144 12 × 12 filters, of which we only display a subset here. Each column contains filters or basis functions of a particular type, and each of the rows has a number relating to which row of the filter or basis function matrix is displayed. (a) PCA ($\mathbf{W}_P$): The first, fifth, seventh etc principal components, calculated from equation (7), showing increasing spatial frequency. There is no need to show basis functions and filters separately here since for PCA, they are the same thing. (b) ZCA ($\mathbf{W}_Z$): The first six entries in this column show the 1-pixel wide centre–surround filter which whitens while preserving the phase spectrum. All are identical, but shifted. The lower six entries (37, 60... 144) show the basis functions instead, which are the columns of the inverse of the $\mathbf{W}_Z$ matrix. (c) **W**: the weights learnt by the ICA network trained on $\mathbf{W}_Z$-whitened data, showing (in descending order) the DC filter, localized oriented filters, and localized checkerboard filters. (d) $\mathbf{W}_I$: The corresponding ICA filters, calculated according to $\mathbf{W}_I = \mathbf{W}\mathbf{W}_Z$, looking like whitened versions of the **W**-filters. (e) **A**: the corresponding basis functions, or columns of $\mathbf{W}_I^{-1}$. These are the patterns which optimally stimulate their corresponding ICA filters, while not stimulating any other ICA filter, so that $\mathbf{W}_I\mathbf{A} = \mathbf{I}$.

*The images (gif files) used are available in the Web directory ftp://ftp.cnl.salk.edu/pub/tony/VRimages.