The "Independent Components" of Natural Scenes are Edge Filters

Anthony J. Bell and Terrence J. Sejnowski
Vision Research, **37(23), 1997**

Markus Koskela

12 October 2000

Introduction

- Neurons in the visual cortex form edge detectors
- Barlow: redundancy reduction, each detector should be as statistically independent from the others as possible
- Second-order statistics (PCA) do not suffice to predict the formation of localized edge detectors
- Sparse or "minimum entropy" coding (ICA): each detector is activated as rarely as possible
- Results:
  - ICA on natural image patches yields localized edge detectors
  - Outputs of ICA filters are more sparsely distributed than those of other decorrelating filters (PCA, ZCA)

Decorrelation and Independence

- Images are transformed with a matrix of filters: $\mathbf{u} = \mathbf{W}\mathbf{x}$
- $\mathbf{W}$ is a decorrelating matrix if the covariance matrix of $\mathbf{u}$ is diagonal
- There are many matrices $\mathbf{W}$ which decorrelate:
- The orthogonal (global) solution $\Rightarrow$ PCA
- The symmetrical (local) solution $\Rightarrow$ ZCA (zero-phase whitening filter)
- The independent (semi-local) solution $\Rightarrow$ ICA
  - ICA on natural images produces decorrelating filters which are sensitive to both locality and frequency information (Gabor filters, wavelets)

The used ICA algorithm

- Unlike PCA and ZCA, an ICA solution may not exist and a given algorithm may not find the solution even if it does exist
- Infomax: maximization by stochastic gradient ascent of the joint entropy of the transformation squashed by a nonlinear function (logistic, tanh)
$$\Delta\mathbf{W} \propto [\mathbf{W}^T]^{-1} + \hat{\mathbf{y}}\mathbf{x}^T$$
- Using the natural (relative) gradient instead of the absolute gradient, the learning rule becomes

$$\Delta\mathbf{W} \propto \frac{\partial H(\mathbf{y})}{\partial \mathbf{W}}\mathbf{W}^T\mathbf{W} = (\mathbf{I} + \hat{\mathbf{y}}\mathbf{u})\mathbf{W}$$

Methods

- 12×12 -sized samples from four natural scenes
- "sphering": $\mathbf{x} \leftarrow 2\mathbf{W}_Z(\{\mathbf{x}\} - \langle\mathbf{x}\rangle)$
  - removes first- and second-order statistics
  - makes the covariance matrix equal to $4\mathbf{I}$
- Initialization of matrix $\mathbf{W}$ to the identity matrix
- Training with infomax
- The full ICA transform is then $\mathbf{W}_I = \mathbf{W}\mathbf{W}_Z$
- Results were verified by using a randomly initialized $\mathbf{W}$ and not prewhitened data

Results

- PCA filters are spatially global and ordered in frequency
- ZCA filters are spatially local and ordered in phase
- ICA filters are semi-local and most of them have a specific orientation
- The distribution (image histogram) for the ICA filters is more sparse (reflected in the kurtosis measure)
- ICA joint histogram of averaged filter pairs capture better the characteristics of the product of sparse univariate distributions (independence criterion)

Noise

- In PCA, noise is associated with high frequency and low amplitude
  - sharp edges contain many high-frequency, low-amplitude components
- In ZCA, removal of noise model results in single-pixel wide peaks
  - photon shot noise
  - photoreceptors and the brain can, however, operate on single detector regime
- In general, there is nothing a priori to distinguish signal from noise ⇒ "noiseless infomax" may be appropriate

Second and higher-order statistics

- Models restricted to second-order statistics are unable to produce local oriented receptive fields
- Second-order statistics correspond to the amplitude spectrum of a signal; higher-order statistics correspond to the phase spectrum
- The phase spectrum can be considered as the informative part of an image: if we remove the phase spectrum, it looks like noise while if we remove amplitude information, the image is still recognizable

Sparseness

- Barlow: The goal of an image transformation should be to convert "higher-order redundancy" to "first-order redundancy

$$H(u_1, u_2) = H(u_1) + H(u_2) - I(u_1, u_2)$$

- Low entropy in $H(u_i)$ can mean that the corresponding distribution is sparse
- Sparse distributions have positive kurtosis (super-gaussian)
- Infomax emphasizes independence over sparseness but the output is still sparse

Biological significance

- V1 simple cell receptive fields are local and oriented (like $\mathbf{W}_I$) but much more complex
- Objections on using $\mathbf{W}_I$ as a model of the retinocortical transformation:
  - filters are mostly of high spatial frequency (unlike the ones in the visual cortex)
  - non-local learning rule in the algorithm
  - non-Hebbian learning rule