Information Maximization in Face Processing

Marian Stewart Bartlett
Institute for Neural Computation
University of California, San Diego
marni@salk.edu

Abstract

This perspective paper explores principles of unsupervised learning and how they relate to face recognition. Dependency coding and information maximization appear to be central principles in neural coding early in the visual system. These principles may be relevant to how we think about higher visual processes such as face recognition as well. The paper first reviews examples of dependency learning in biological vision, along with principles of optimal information transfer and information maximization. Next, we examine algorithms for face recognition by computer from a perspective of information maximization. The eigenface approach can be considered an unsupervised system that learns the first and second-order dependencies among face image pixels. Eigenfaces maximizes information transfer only in the case where the input distributions are Gaussian. Independent component analysis (ICA) learns high-order dependencies in addition to first and second-order relations, and maximizes information transfer for a more general set of input distributions. Face representations based on ICA gave better recognition performance than eigenfaces, supporting the theory that information maximization is a good strategy for high level visual functions such as face recognition. Finally, we review perceptual studies suggesting that dependency learning is relevant to human face perception as well, and present an information maximization account of perceptual effects such as the atypicality bias, and face adaptation aftereffects.

Introduction

Horace Barlow has argued that statistical dependencies in the sensory input contain structural information about the environment, and that a general strategy for sensory systems is to learn the expected dependencies. Figure 1 is an example from Barlow’s 1989 paper illustrating the idea that the percept of structure is driven by the dependencies in the sensory input. The set of points on the left was randomly selected from a Gaussian distribution. In the figure on the right, half of the points were generated from a Gaussian distribution and the other half were generated by rotating those points 5 degrees about the centroid of the distribution. The dependence between pairs of dots gives this image a percept of structure.

Figure 1. Example of visual structure from statistical dependencies. (From Barlow, 1989).
**Adaptation.** Adaptation mechanisms are examples of encoding dependencies in early visual processing. See Figure 2. As pointed out by Barlow (1989), a first order redundancy is mean luminance. Adaptation mechanisms take advantage of this redundancy by expressing lightness values relative to the mean. In this example, these two squares have the same gray value, but they are embedded in regions with different mean luminance, and the perception of brightness is different. A second order redundancy is luminance variance, which is a measure of contrast. Contrast gain control mechanisms express contrast relative to the expected contrast. Here is an example of contrast gain control. These two squares have the same contrast, but this one appears to have higher contrast when it is embedded in a region of low mean contrast.

![Luminance and Contrast Adaptation](image)

**Figure 2.** Simultaneous contrast. Top: Removing first order redundancy in the luminance signal. The inner squares have the same gray value. The percept of brightness is relative to the mean brightness in the region. Bottom: Removing second-order redundancy in the luminance signal. The inner squares have the same contrast. The percept of contrast is relative to the mean contrast in the region.

**Information maximization in neural codes.** A more comprehensive way to encode the redundancy in the input signal, instead of just learning the means, is to learn the probability distribution of signals in the environment. Learning the redundancy means learning which values occur most frequently. Neurons with a limited dynamic range can increase the information that the response gives about the signal by placing the more steeply sloped portions of the transfer function in the regions of highest density, and shallower slopes at regions of low density. What we know from information theory is that the optimal response function for maximizing information transfer is to match the transfer function to the cumulative probability density of the input signal. See Figure 3.

![Signal distribution, Transfer function, Response distribution](image)

**Figure 3.** The response distribution is obtained by multiplying the input distribution by the transfer function. When the transfer function matches the cumulative probability density of the input, the output distribution is uniform and information transfer is maximized.
Simon Laughlin was one of the first people to explore this issue in neural response functions, and he coined the term ‘infomax’. Figure 4a shows an estimate of the probability of contrasts measured in the fly environment. Figure 4b shows the cumulative probability density of contrast in the fly’s environment, and it is the response function that is predicted if the fly’s visual system is performing information maximization. Superimposed is the observed transfer function. Green dots show the observed response of the fly LMC cells as a function of contrast. There is a close match between the observed transfer function and that predicted by information maximization.

![Distribution of contrast in environment](image1) ![Response function of LMC cells](image2)


There are numerous other examples of information maximization in neural codes (See Simoncelli and Olshausen, 2001, for a review). For example, Atick and Redlich (1991) showed this for contrast sensitivity as a function of spatial frequency in both Macaque retinal ganglion cells and psychophysical contrast sensitivity functions. Macleod and von der Twer have shown that information maximization of chromatic signals also account for spectral sensitivities in the primate color opponent system as well. Infomax principles have also been demonstrated in the temporal frequency response in cat LGN (Dan et al., 1996), primary visual cortex (Vinge & Gallant, 2000). It has also been shown that short-term adaptation reduces dependencies between neurons both physiologically (e.g. Carandini, 1998) and using psychophysics (e.g. Webster, 1996).

**Information theory**: Here we review some basic concepts from information theory that are behind the information maximization principle.

\[ I(x) = -\log(P(x)) \]  
(1)

The information I in a signal x is inversely proportional to its probability. Low probability events contain more information. For example, if you tell me that you spoke to a student in your class who was 20 years old, that would not tell me much about who it was, but if you tell me you spoke to a student who was 43, that would give me more information about who it was.

\[ H(x) = E[I(x)] \]  
(2)
The entropy $H$ is the expected value of the information over the whole probability distribution. Entropy is maximized by flat distributions and minimized by peaked distributions.

The mutual information between an input $x$ and an output $y$, $I(x,y)$, is the expected value of the joint probability of $x$ and $y$, but it can be expressed as the entropy of $Y$ minus the entropy of $y$ given $x$.

$$I(x,y) = H(y) - H(y|x) \quad (3)$$

The entropy of $y$ given $x$ is the noise, since it is the information in the signal that is not accounted for by the input. Therefore, maximizing $I(x,y)$ is equivalent to maximizing $H(y)$. In other words, maximizing information transfer is equivalent to maximizing the entropy of the output.

**Hebbian Learning.** These concepts relate to learning at the neuron level. Hebbian learning is a model for long-term potentiation in neurons, in which weights are increased when the input and output are simultaneously active. The weight update rule is typically formulated as the product of the input and the output activations. This simple rule learns the covariance structure of the input (i.e. the second-order relations). In the case of a single input and output unit, Hebbian learning maximizes the information transfer between the input and the output (Linkser, 1988). Hence long term potentiation is related to information maximization. For multiple inputs and outputs, however, Hebbian learning doesn’t maximize information transfer except in the special case where all of the signal distributions are Gaussian. Many types of natural signals have been shown to have non-Gaussian distributions, in which the distributions are much more steeply peaked (e.g. Lewicki, 2002).

**Independent component analysis.** The independent component analysis learning rule developed by Bell & Sejnowski (1995) is a generalization of Linsker’s information maximization principle to the multiple unit case, $X=(x_1 \ldots x_n)$, $Y=(y_1 \ldots y_n)$. In this case, the information transfer between the input $X$ and output $Y$ is maximized by maximizing the joint entropy of the output, $H(Y)$.

Maximizing $H(Y)$ can minimize the mutual information between the individual outputs $Y=(y_1 \ldots y_n)$. This can be seen as follows: The equation for the joint entropy of the output $Y$ is the sum of the individual entropies minus the mutual information between them:

$$H(y_1,y_2) = H(y_1)+H(y_2)-I(y_1,y_2). \quad (4)$$

Maximizing $H(y_1,y_2)$ encourages $I(y_1,y_2)$ to be small. The mutual information is guaranteed to be minimized when each transfer function matches the cumulative probability density of the corresponding source (Parga). So the take-home message is that information maximization and redundancy reduction are highly related.

The independent component analysis learning rule has a Hebbian learning term, but instead of being a straight product between the input and the output activations, the Hebbian learning is between the input and the gradient of the output:

$$\Delta W = \alpha [(W^T)^{-1} + y'x^T] \quad (5)$$
where $W$ is a weight matrix and $\alpha$ is the learning rate.\(^1\)

**Sparse codes.** For maximizing information with a single unit, the optimal output distribution is flat or uniform, but when maximizing information in a population code, the output distribution of the individual units is different. ICA maximizes the joint entropy of the output, but minimizes the marginal entropies (the entropy of the individual outputs). So in a population code, infomax makes the joint output distribution flat, but the distributions of the individual outputs are highly peaked, or sparse. See Figure 5.

![Figure 5](image)

Figure 5. In a population code, a maximum entropy distribution (solid line), can be created from individual neurons with sparse distributions (dashed lines).

This falls out of the equation for joint entropy, which is repeated here. The independent component solution to equation 4 which minimizes $I(y_1,y_2)$ also minimizes $H(y_1)$ and $H(y_2)$. (We hold $H(y_1,y_2)$ constant and minimize $I(y_1,y_2)$. This in turn minimizes $H(y_1)$ and $H(y_2)$.)

$$H(y_1,y_2) = H(y_1) + H(y_2) - I(y_1,y_2).$$

Barlow (1989) advocated minimum entropy coding for this reason, where the individual neurons have minimum entropy, or sparse distributions. It can be shown that maximizing sparseness without information loss is equivalent to ICA and infomax (Barlow; Olshaussen & Field). (Again we hold $H(y_1,y_2)$ constant, and this time minimize $H(y_1)$ and $H(y_2)$, which in turn minimizes $I(y_1,y_2)$.)

Neurons early in the visual system may do infomax within a neural response, whereas neurons higher in the visual system may do infomax in population codes. Certainly there is a trend for neural responses to become more sparse higher in the visual system (e.g. Rolls).

**Relationships of ICA to V1 receptive fields.** A number of relationships have been shown between independent component analysis and the response properties of visual cortical neurons. For example, applying ICA to a set of natural images produces a set of learned weights that are local, spatially opponent edge filters (Bell & Sejnowski, 1996; Olshausen & Field, 1996).

Conversely, passing a set of natural scenes through a bank of Gabor filters produces outputs that are at least pairwise independent when you also include gain control mechanisms such as the ones proposed by David Heeger to model V1 responses (Field, 1991; Simoncelli, 1997). When ICA is

\[^{1}\] Another version of this learning rule employs the *natural gradient*, which is the gradient multiplied by $W^TW$ (Bell & Sejnowski, 1996), which has important regularization properties. Here the Hebbian learning term is between the input and the *natural gradient* of the output.

$$\Delta W = \alpha (I + y^T \chi^T W^T) W.$$
applied to chromatic natural scenes, the set of weights segments into color-opponent and broadband filters, where the color-opponent filters code for red-green and blue yellow (Wachtler, Lee & Sejnowski, 2001). Moreover, a two-layer ICA model learned translation invariant receptive fields related to complex cell responses (Hyvarinen & Hoyer, 2001).

**Application to face recognition by computer**

These learning principles have been applied to face images for face recognition. Eigenfaces (Turk & Pentland, 1991; Cottrell & Metcalfe, 1991) is essentially an unsupervised learning strategy that learns the second-order dependencies among the image pixels. It applies principal component analysis (PCA) to a set of face images. Principal component solutions can be learned in neural networks with simple Hebbian learning rules (e.g. Oja 1989). Hence one way to interpret Eigenfaces, albeit not the way it is usually presented in the computer vision literature, is that it applies a Hebbian learning rule to a set of image pixels.

![Figure 6. Eigenfaces. a: Example face space in which the first 3 pixel values of 15 face images is plotted. The first two eigenvectors are depicted by dashed lines. b: 4 eigenvectors of a 60x90 pixel face space.](image)

Each image is a point in a high dimensional space defined by the grayvalue taken at each of the pixel locations. See Figure 6. Principal component analysis rotates the axes to point in the directions of maximum variance of the data. The faces are recoded by their coordinates with respect to the new axes. These coordinates are the \( a_i \) in Figure 6. The axes are defined by the eigenvectors of the covariance matrix. These eigenimages form a basis set where each face is a weighted combination of the basis images. The new coordinates, or weights, form a representation on which identity classification is performed.

This approach took the computer vision community by surprise in the early 1990’s. It performed considerably better than contemporary approaches that focused on measuring specific facial features and the geometric relationships between them. The approach seemed counterintuitive many in computer vision. Why should this representation give a good description of face images? The success of the approach may have been due to the fundamental strategy of learning the dependencies in a population of face images.

Principal component analysis learns the second-order dependencies among the pixels but does not learn high order dependencies. For example, if we change the pixels so that the covariances stay the same, then PCA gives the same answer.

Bartlett, Movellan, & Sejnowski (2002) developed a representation for face recognition based on independent component analysis. Independent component analysis learns the high order
dependencies in addition to the ones that PCA learns. We compared it to eigenfaces to see if encoding more of the dependencies would result in better recognition performance.

**Second order vs. high-order dependencies in face images.** The second-order dependencies correspond to the amplitude spectrum of an image, and the high order dependencies correspond to the phase. More precisely, a theorem by Einstein and Wiener showed that the Fourier transform of the covariance matrix of an image it corresponds to the amplitude spectrum of the image. What is left over is the phase.

Figure 7 illustrates the contribution of high order dependencies to face perception. Experiments such as those by Oppenheim and Lim suggest that the phase spectrum contains the structural information that drives drives human perception. If you do a Fourier transform on the face images in Figure 7, and then reconstruct a face using the amplitude spectrum of the male but the phase spectrum of the female, the face looks like the female. Likewise, if we use his phase spectrum and her power spectrum the image looks like him.

Bartlett, Movellan, & Sejnowski described two ways to apply ICA to face images. One way is similar to the PCA example illustrated above. Each image is an observation in a high dimensional space where the dimensions are the pixels. (See Figure 8a). It is also possible to treat the images as observations and the pixels as the variables. Each pixel is plotted according to the grayvalue it takes on over a set of face images. (See Figure 8b).
The ICA model decomposes images as $X = AS$ where $A$ is a mixing matrix and $S$ is a matrix of independent sources. In Architecture I, $S$ contains the basis images, and $A$ contains the coefficients. Whereas in Architecture II, $S$ contains the coefficients and $A$ contains the basis images. The basis images learned in Architecture I are spatially local, whereas the basis images learned in Architecture II are global, or configurational. (See Figure 9).

Figure 9. Image synthesis model for the two architectures. The ICA model decomposes images as $X = AS$, where $A$ is a mixing matrix and $S$ is a matrix of independent sources. In Architecture I (top), the independent sources $S$ are basis images and $A$ contains the coefficients, whereas in Architecture II (bottom) the independent sources $S$ are face codes (the coefficients) and $A$ contains the basis images.

Bartlett, Movellan and Sejnowski (2003) compared face recognition performance of ICA to PCA on a set of FERET face images. This image set contained 425 individuals with up to four images each: Same day with a change of expression, a different day up to 2 years later, and a different day with a change of expression. Recognition was done by nearest neighbor using cosines as the similarity measure. The results are shown in Figure 10. Both ICA face representations outperformed PCA for the pictures taken on a different day, which are the much harder recognition conditions.
Figure 10. Face recognition performance on the FERET database. ICA architectures I and II are compared to PCA (eigenfaces). Gray bars show improvement in performance following class-specific subselction of basis dimensions.

The ICA representations maximize information transmission in the presence of noise and thus they may be more robust to sources of noise such as variations in lighting conditions, facial expressions, and changes in hair and make-up.

When subsets of bases are selected, ICA and PCA define different subspaces. Bartlett, Movellan, & Sejnowski examined face recognition performance following subspace selection with both ICA and PCA. Bases were selected by class discriminability, which we defined as the ratio of the variance within faces to the variance between faces. The gray extensions in Figure 10 show the improvement by subspace selection. ICA defined subspaces encoded more information about facial identity than PCA-defined subspaces. We also explored subselction of PCA bases ordered by eigenvalue. ICA continued to outperform PCA. Moreover, combining the outputs of the two ICA representations gave better performance then either one alone. The two representations appeared to capture different information about the face images.

These findings suggest that information maximization principles are an effective coding strategy for face recognition by computer. Namely, the more dependencies that were encoded, the better the recognition performance.

Local representations versus factorial codes. Draper and colleagues (2003) conducted a comparison of ICA and PCA on a substantially larger set of FERET face images consisting of 1196 individuals. This study supported the finding that ICA outperformed PCA, and included a change in lighting condition which we had not previously tested. ICA with architecture II obtained 51% accuracy on 192 probes with changes in lighting, compared to the best PCA performance of 40% correct.

An interesting finding to emerge from the Draper study is that the ICA representation with Architecture II outperformed Architecture I for identity recognition. See Figure 11. According to arguments by Barlow, Atick, and Field (Barlow, 1989; Atick, 1992a; Field, 1994), the sparse, factorial properties of the representation in Architecture II should be optimal for face coding. Architecture II provides a factorial face code, in that each element of the face representation is independent of the others (i.e. the coefficients are independent). Although the previous study showed no significant difference in recognition performance for the two architectures, there may have been insufficient training data for a difference to emerge. Bartlett (2001) predicted that The ICA-factorial representation may prove to have a greater advantage given a much larger training set of images. Indeed, this prediction was born out.

When the task was changed to recognition of facial expressions, however, Draper et al found that the ICA representation from Architecture I outperformed the ICA representation from Architecture II. The task was to recognize 6 facial actions, which are individual facial movements approximately corresponding to individual facial muscles. Draper et al attributed their pattern of results to differences in local versus global processing requirements of the two tasks. Architecture I defines local face features whereas Architecture II defines more configural face features. A large body of literature in human face processing points to the importance of configural information for identity recognition, whereas the facial expression recognition task in this study may have greater emphasis on local information. This speaks
to the issue of separate basis sets for expression and identity. There is evidence in functional neuroscience for separate processing of identity and expression in the brain (e.g. Haxby, 2000.) Here we obtain better recognition performance when we define different basis sets in which we switch what is treated as an observation versus what is treated as an independent variable for the purposes of information maximization.

![Image](image.png)

**Figure 11.** Face recognition performance with a larger image set from Draper et al. (2003).

### Dependency learning and face perception

There are a number of perceptual studies that support the relevance of dependency encoding to human face perception. A large body of work showed that unsupervised learning of 2nd order dependencies successfully models a number of aspects of human face perception including similarity, typicality, recognition accuracy, and other-race effects (e.g. O’Toole et al., 1994; Hancock et al., 1998; Calder). Moreover, one study found that ICA better accounts for human judgments of facial similarity than PCA, supporting the idea that the more dependencies are encoded, the better the model of human perception for some tasks (Hancock, 2000). There is also support from neurophysiology for information maximization principles in face coding. The response distributions of IT face cells are sparse and there is very little redundancy between cells (Rolls, 2004; Rolls & Tovee, 1995).

A perceptual discrimination study by Parraga, Troscianko, and Tollhurst (2000) supported information maximization in object perception. Sensitivity to small perturbations from morphing was highest for pictures with a natural second-order statistics, and degraded as the second order statistics were made less natural.

Adaptation aftereffects in face perception have also been demonstrated (e.g. Webster, 1999; Leopold, O’Toole, Vetter, & Blantz, 2001). For example, after adapting to a distorted face, a neutral face appears distorted in the opposite direction. This is a high-level version of the adaptation processes discussed earlier in this paper. Similar effects have been shown for race, gender, and facial expressions (Kaping et al., 2002). Hence it appears that faces may be coded according to expected values along these dimensions. Interestingly, adaptation to a nondistorted face does not make the distorted face appear more distorted. This is consistent with an information maximization account of adaptation since adapting to a neutral face would not shift the population mean.
Tanaka and colleagues showed a related effect with typical and atypical faces, where instead of adaptation, it was life experience that was shaping the probability distributions of faces (Tanaka, Giles, Kreeman, & Simon, 1998). The Tanaka study showed that morphs between typical and atypical parents appear to be more similar to the atypical parent. Figure 12 shows an example 50/50 morph between typical and atypical faces. In a 2-alternative forced choice, subjects chose the atypical parent as more similar about 60% of the time.

![Sample morph between typical and atypical parent faces](image)

Figure 12. Sample morph between typical and atypical parent faces, from Tanaka, Giles, Kreeman, & Simon, (1998).

Tanaka et al suggested an attractor network account for this effect, where the atypical parent has a larger basin of attraction than the typical parent. See Figure XX. Bartlett & Tanaka implemented an attractor network model of face perception, and indeed successfully modeled the atypicality bias. Inspection of this model provides further insights about the potential role of redundancy reduction in face perception.

In this model, a decorrelation step is necessary in order to encode highly similar patterns in a Hopfield attractor network. Decorrelation was necessary in order for each face to take on a distinct pattern of sustained activity. This Hence the atypicality bias arose from a decorrelation process inherent to producing separate internal representations for highly similar faces.

An alternative account of the atypicality bias is provided by information maximization and redundancy reduction. This account is compatible with the attractor network hypothesis but more parsimonious. Figure 13a shows the probability distribution of face property X such as nose size. A typical face has a value near the mean for this property, and an atypical face has a value that is in the tail. Figure 13b shows the transfer function between signal value and percept. Let us suppose it is adapted to match the cumulative probability density so that it performs information maximization. In this case, typical faces fall on a highly sloping region of the transfer function, whereas atypical faces fall on a flatter region. The 50% morph is mapped closer to the atypical face because of the shape of the transfer function. The infomax account makes an additional prediction. Although it is well known that faces rated as ‘atypical’ tend to also be easier to recognize, this model predicts that subjects will be less sensitive to physical changes in the physical properties of atypical faces.

Perceptual effects such as the other-race effect fall out of the same information maximization model. In Figure 13, simply replace “typical” and “atypical” with “same race” and “other race.” Since subjects have more experience with same-race faces, the physical properties of same-race faces would fall near the center of the probability density function. The infomax transfer function would be steeper near same-race stimulus properties, and shallower near ‘other race’ stimulus properties. Walker and Tanaka (in press) recently showed that discrimination performance is superior for same-race than other-race faces, which is consistent with information maximization coding.
Figure 13. Information maximization account of the atypicality bias in the perception of morphed faces. a. Example probability distribution for a face property such as nose size, in which there is higher probability density near typical faces. b. Perceptual transfer function predicted by information maximization (the cumulative density of the distribution on the left). The percept of the 50% morph is mapped closer to the atypical face because of the shape of the transfer function. (See dashed lines).

**Information maximization account of adaptation aftereffects.** This account also applies to face adaptation aftereffects. The neural mechanisms may not be the same, they operate over different time scales, but the information maximization principle may apply to both cases. Consider, for example, that the adapting face is a distorted face in the tail of the face distribution for face feature X such as nose size. See Figure 14a. After adapting, the short-term estimate of probability has increased density near the adapting stimulus, as shown by the dashed curve. Figure 14b shows the transfer functions predicted by information maximization for both pre- and post-adaptation probability densities. These are the cumulative probability density functions. Note that after adaptation, the neutral face is mapped to a distortion in the opposite direction of the adapting stimulus, which matches the psychophysical findings (e.g. Webster 1999). The increased slope near the adapting stimulus also predicts increased sensitivity near the adapting stimulus.

An alternative possibility is that neural mechanisms do not implement unconstrained optimization such as Figure 14b, but that they do constrained optimization the with a family of functions such as a sigmoid. This possibility is illustrated in Figure 15a. Here a sigmoid was fitted to the cumulative probability density of the adaptation distribution using logistic regression. This model also predicts repulsion of the neutral face away from the adapting stimulus. These two models, however, give differing predictions on sensitivity changes, shown in Figure 15b. Sensitivity predictions were obtained by taking the first derivative of the transfer function. Unconstrained information maximization predicts increases in sensitivity near the adapting stimulus. The sigmoid fit to the optimal transfer function predicts a much smaller increase in sensitivity near the adapting stimulus, plus an overall shift in sensitivity towards the adapting stimulus.
Figure 14. a. Example probability density of a face property such as nose size before and after adaptation. Adaptation to a face with large noses changes the short-term probability density of nose size. Extra density is added near the adapting stimulus, and density decreases elsewhere since it must sum to one. B. Perceptual transfer function predicted by information maximization. (the cumulative pdf of the density following adaptation).

Figure 15. a: approximation of the optimal transfer function following adaptation with a sigmoid. b. Sensitivity functions predicted by the two transfer functions shown in Figure XX and XX. Sensitivity predictions were obtained by taking the first derivative of the transfer function. Information maximization (red --) predicts increased sensitivity near the adapting stimulus. Alternatively, a sigmoid fit to the optimal transfer function (green -..) predicts a shift in the sensitivity curve towards the adapting stimulus.

**Infomax in the spatiotemporal domain.** Carrying these learning strategies into the spatiotemporal domain can help learn visual invariances. Viewpoint invariant representations can be obtained by learning temporal relationships between images as an object moves in the environment. There are a number of models of learning invariances from spatiotemporal dependencies (e.g. Foldiak, 1991; O’Reilly, 1992; Becker 1999, Wallis & Rolls, 1997; Stone,
1996; Bartlett & Sejnowski, 1998). For example, Bartlett & Sejnowski (1998) modeled the development of representations that are independent of head pose by learning spatiotemporal relationships. Moreover, spatio-temporal ICA on movies of natural images results in space/time varying filters such as those in MT (Dan).

The long open time of the NMDA channel provides a possible mechanism for learning spatiotemporal relationships in long-term potentiation (O’Reilly, 1992). There is evidence that cells in the primate anterior temporal lobe can learn temporal relationships between visual inputs (Miyashita, 1988). After monkeys were presented a sequence of fractal patterns for a number of weeks, the AIT cells developed tuning curves for the position of the stimuli in the sequence. Responses to neighboring stimuli were correlated, and correlation reduced as the distance in the sequence increased.

**Summary**

Dependency coding and information maximization appear to be central principles in neural coding early in the visual system. Neural systems with limited dynamic range can increase the information that the response gives about the signal by placing the more steeply sloped portions of the transfer function in the regions of highest density, and shallower slopes at regions of low density. The function that maximizes information transfer is the one that matches the cumulative probability density of the input. There is a large body of evidence that neural codes in vision and other sensory modalities match the statistical structure of the environment, and hence maximize information about environmental signals to a degree. See (Simoncelli, 2001) for a review. This paper described how these principles may be relevant to how we think about higher visual processes such as face recognition as well.

Here we examined algorithms for face recognition by computer from a perspective of information maximization. Principal component solutions can be learned in neural networks with simple Hebbian learning rules (Oja, 1989). Hence the Eigenface approach can be considered a form of Hebbian learning model, which performs information maximization under restricted conditions. Independent component analysis performs information maximization under more general conditions. The learning rule contains a Hebbian learning term, but it is between the input and the gradient of the output. Face representations derived from ICA gave better recognition performance than face representations based on PCA. This suggests that information maximization in early processing is an effective strategy for face recognition by computer.

Next, we presented an information maximization account of perceptual effects including other-race effects and typicality effects, and showed how face adaptation aftereffects (e.g. Kaping, 2002; Ng, 2003; Webster, 1999) are consistent with information maximization on short time scales. Two models of information maximization in adaptation were presented, one in which the visual system learns a high-order transfer functions that match the curves in the cumulative probability density, and another in which the cumulative probability density is approximated with the closest fitting sigmoid. These two models gave different predictions for sensitivity post-adaptation. The extent to which information maximization principles account for perceptual learning and adaptation aftereffects in human face perception is an open question and an area of active research.
References


