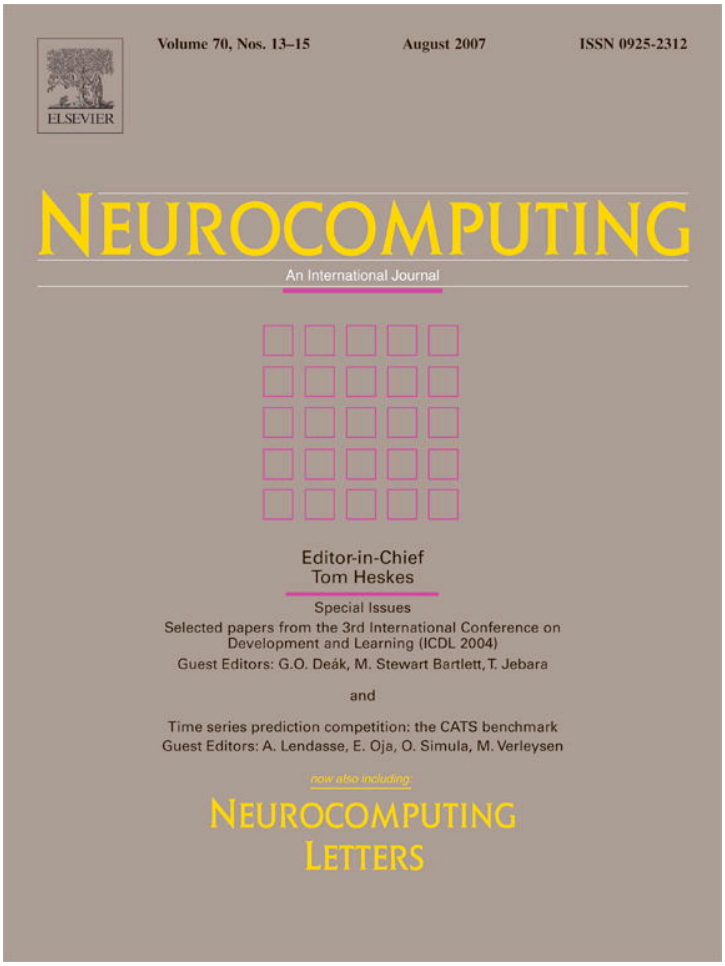

Available online at www.sciencedirect.com
ScienceDirect
Neurocomputing 70 (2007) 2204–2217
NEUROCOMPUTING
www.elsevier.com/locate/neucom

Information maximization in face processing

Marian Stewart Bartlett*

Institute for Neural Computation, University of California, San Diego, USA

Received 5 February 2005; received in revised form 16 August 2005; accepted 1 February 2006
Available online 2 January 2007

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 as an unsupervised system that learns the first- and second-order dependencies among face image pixels. Eigenfaces maximize 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.

Keywords: Face recognition; Information maximization; Independent component analysis; Vision; Learning; Unsupervised learning; Machine learning; Visual learning

1. Introduction

A line of research lead by Attneave and Barlow [1,3,7] argues 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. Fig. 1 illustrates the idea that the percept of structure is driven by the dependencies in the sensory input. The set of points in 1c was randomly selected from a Gaussian distribution. In 1d, half of the points were generated from a Gaussian distribution and the other half were generated by rotating those points 5° about the centroid of the distribution. The dependence between pairs of dots gives the image in 1d a percept of structure.

1.1. Examples of dependency learning in the visual system

Adaptation mechanisms are examples of encoding dependencies in early visual processing. As pointed out in [7], a first-order redundancy is mean luminance. Adaptation mechanisms take advantage of this redundancy by expressing lightness values relative to the mean. Fig. 2 illustrates this idea. The two squares in the top row have the same grayvalue, but they are embedded in regions with different mean luminances. The square on the left has higher luminance than the local mean, and is perceived to be brighter than the square on the right. A second-order redundancy is luminance *variance*, which is a measure of contrast. Contrast gain control mechanisms express contrast relative to the expected contrast. The bottom row of Fig. 2 illustrates the dependence of perceived contrast on the local mean contrast. The two inner squares have the same physical contrast, but the one on the right is perceived to have higher contrast when it is embedded in a region of low mean contrast.

*Tel.: +1 858 822 5241; fax: +1 858 534 2014.
E-mail address: marni@salk.edu.

0925-2312/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2006.02.025

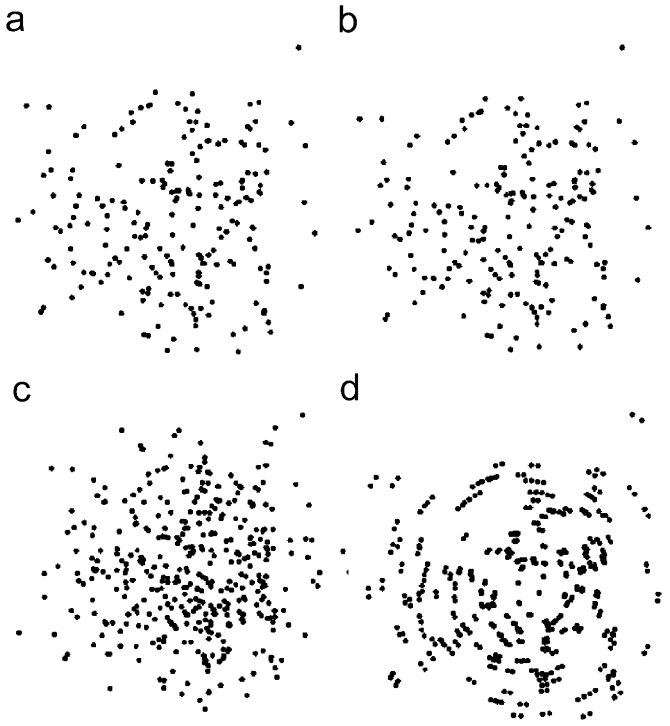

Fig. 1. Example of visual structure from statistical dependencies, adopted from [5]. (a) A set of 200 points randomly selected from a Gaussian distribution. (b) The points in a, rotated by 5°. (c) 400 randomly selected points. (d) The union of the points in a with the points in b.

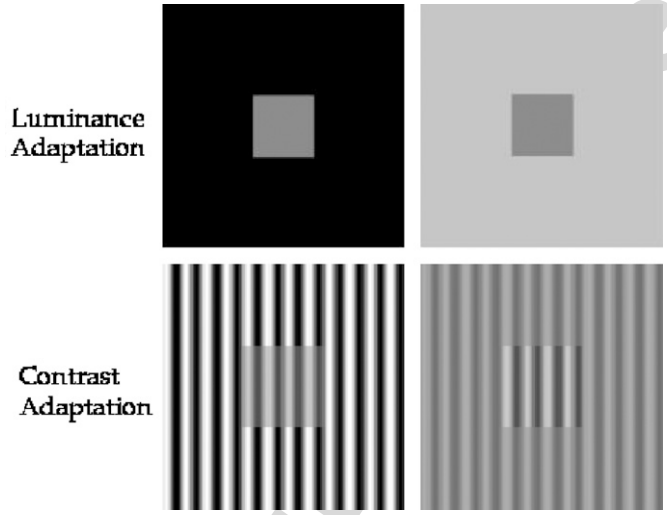

Fig. 2. First- and second-order redundancies in the luminance signal. Top: The inner squares have the same grayvalue. The percept of brightness is relative to the mean brightness in the region. Bottom: The inner squares have the same contrast. The percept of contrast is relative to the mean contrast in the region.

1.2. 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 where 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. When the transfer function matches the cumulative probability density of the input, the output distribution is uniform. See Fig. 3. The probability distribution of the output is determined by the probability distribution of the input signal and the slope of the transfer function. Specifically, for monotonically increasing transfer functions, the probability of the output is obtained by dividing the probability of the input by the slope of the transfer function [57].

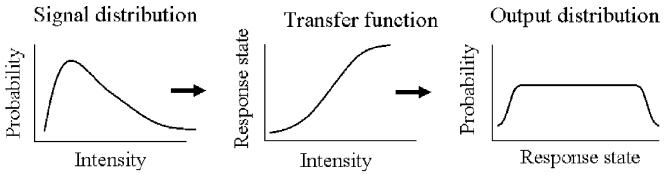

Fig. 3. When the transfer function matches the cumulative probability density of the input, the output distribution is uniform, which maximizes information transfer in the case of limited response range. The output distribution is the probability distribution of different response levels of the neuron.

Laughlin [38] was one of the first papers to explore this issue in neural-response functions. Laughlin measured the contrasts in the environment of the blowfly, and estimated a probability density from these measures. He next estimated the response function of the LMC cells in the blowfly compound eye by measuring the response at a range of contrasts. As shown in Fig. 4, the LMC responses were a close fit to the cumulative probability density of contrast in the fly's environment, which is the response function that is predicted if the fly's visual system is performing information maximization under conditions of limited dynamic range.

There are numerous other examples of information maximization in neural codes (see [65] for a review). For example, Atick and Redlich [2] showed that contrast sensitivity functions increase information transfer for low and middle spatial frequencies in natural scenes. They showed that when contrast sensitivity functions are multiplied by $1/f$, which is the amplitude observed for natural scenes at frequency f, the output amplitude spectrum is flat through about 3 cycles/degree. This whitening of the retinal ganglion cell amplitude spectrum can be interpreted as reducing the second-order dependencies at low spatial frequencies, since the amplitude spectrum is a measure of second-order statistics. The relationship between redundancy reduction and information maximization is discussed in more detail below. Information maximization principles have also been demonstrated in the temporal frequency response in cat LGN [23], and primary visual

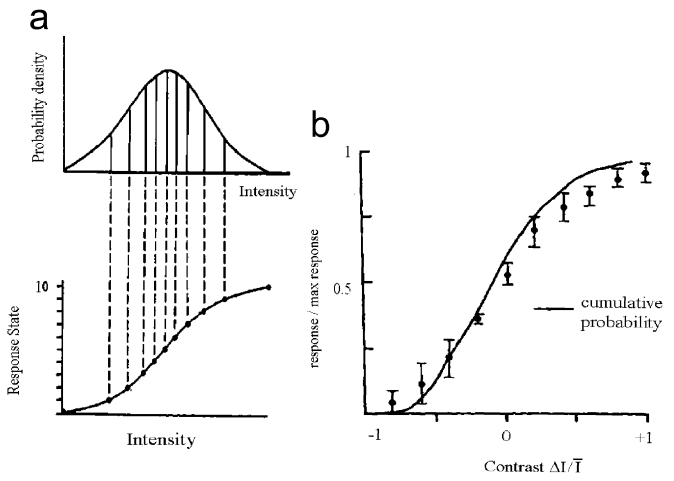

Fig. 4. Information maximization in blowfly compound eye, reprinted with permission from [38]. (a) "The coding strategy for maximizing a neurons information capacity by ensuring that all response levels are used with equal frequency. Upper curve—probability density function for stimulus intensities. Lower curve—the intensity–response function that implements the strategy. In this example the neuron has 10 response states, corresponding to 10 "just noticeable differences" in response. The intensity–response function insures that the interval between each response level encompasses an equal area under the intensity distribution, so that each state is used with equal frequency. In the limit where the states are vanishingly small this intensity–response function corresponds to the cumulative probability function for stimulus intensities. (b) The contrast–response function of light adapted fly LMC's (large monopolar cells) compared to the cumulative probability function for natural contrasts…". Reprinted with permission from Verlag Z. Naturforsch.

cortex [73]. Von der Twer and Macleod [74] have shown that related principles apply to spectral sensitivities in the primate color opponent system. We will return to von der Twer and Macleod's model in the discussion. It has also been shown that short-term adaptation reduces dependencies between neurons both physiologically (e.g. [18]) and using psychophysics (e.g. [80]).

1.3. Overview

Dependency coding and information maximization appear to be central principles in neural coding early in the visual system. This paper describes how these principles may be relevant to how we think about higher visual processes such as face recognition as well. Section 2 reviews some concepts from information theory. Section 3 describes how information maximization is achieved in two unsupervised learning rules, Hebbian learning and independent component analysis (ICA).

Section 4 examines 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 [53]. Hence the eigenface approach can be considered a form of Hebbian learning model, which performs information maximization under restricted conditions. ICA 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 principal component analysis (PCA). This suggests that information maximization in early processing is an effective strategy for face recognition by computer.

Section 5 presents an information maximization account of perceptual effects including other-race and typicality effects, and shows how face adaptation aftereffects (e.g. [36,51,81]) are consistent with information maximization on short time scales. Two models of information maximization in adaptation are 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 give different predictions for sensitivity post-adaptation.

2. Some concepts from information theory

Here we review some basic concepts from information theory that are behind the information maximization principle. Shannon [64] defined the information I in a message x as inversely proportional to its probability:

$$I(x) = -\log_2(P(x)). \quad (1)$$

The formula is based on the number of bits used to transmit a message x in a digital system, where $I(x)$ is the length of the code in bits. Shannon's optimal coding theorem states that the code length defined in Eq. (1) has the smallest expected length over the full set of messages. The intuition is that we should use shorter length codes to represent commonly occurring signal elements, and it shows the fundamental relationship between information theory and the principle of minimum description length [60].

In Shannon's formulation, the code length for event x under the optimal coding scheme (Eq. (1)) provides a measure of the information provided by event x. Information is the amount of surprise, where low-probability events contain more information than high-probability events. An intuitive example of the inverse relationship between information and probability is the professor who told the TA he spoke to a student in the class who was 20 years old. In this case age gave very little information about which student it was, but had the age been 43 it would have provided much more information about who it was.

The expected value of the information over the whole probability distribution is the entropy, H:

$$H(x) = E[I(x)]. \quad (2)$$

Entropy is a measure of the uncertainty of the distribution. Highly peaked distributions have low uncertainty—the same outcome is observed almost all of the

time, whereas spread-out distributions have high uncertainty. In the case of limited dynamic range, entropy is maximized by a uniform distribution across that dynamic range. Thus histogram equalization is a form of entropy maximization.

The mutual information $I(x, y)$ between two signals x and y is the relative entropy between the joint distribution and the marginal distributions: $I(x, y) = H(x) + H(y) - H(x, y)$. If we consider a function $y = f(x)$, then $I(x, y)$ is the mutual information between the input and the output of this function. The entropy of the output, $H(y)$, is the sum of the uncertainty in y that is explained by x, $I(x,y)$, and the uncertainty in y that is not explained by x, $H(y|x)$. Thus $I(x, y)$ can also be expressed as follows [22]:

$$I(x, y) = H(y) - H(y|x). \tag{3}$$

Eq. (3) is the Shannon information transfer rate. The term $H(y|x)$ is noise, since it is the information in the output that is not accounted for by the input, and it does not depend on the transfer function f. Therefore, finding a function $y = f(x)$ that maximizes $I(x, y)$ is equivalent to finding a function that maximizes $H(y)$ [44]. In other words, maximizing the entropy of the output also maximizes information transfer.

3. Information maximization and unsupervised learning rules

3.1. 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 adjusted in proportion to the input and output activities. The weight update rule is typically formulated as the product of the input and the output activations. Because simultaneously active inputs cooperate to produce activity in an output unit, Hebbian learning finds the correlational structure in the input. See [13] for a review.

For a single output unit with a Gaussian distribution, Hebbian learning maximizes the information transfer between the input and the output [44]. The explanation is outlined as follows. It can be shown that Hebbian learning maximizes the activity variance of the output subject to saturation bounds. For a single output unit with a Gaussian distribution, the Shannon information transfer rate (Eq. (3)) is maximized by maximizing the variance of the output. This is because maximizing the variance of a Gaussian distribution maximizes its entropy, and as seen above, maximizing the entropy of the output maximizes $I(x, y)$ since $H(y|x)$ is noise.

Hence long-term potentiation is related to information maximization. Linsker [44] argued for maximum information preservation as an organizational principle for a layered perceptual system. For a code that spans multiple output units, however, Hebbian learning does not 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 nonGaussian distributions, in which the distributions are much more steeply peaked [24,27,33,41].

3.2. ICA

The ICA learning rule developed by Bell and Sejnowski [16] 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, $H(Y)$. As discussed above (Eq. (3)), finding a function $Y = f(X)$ that maximizes $I(X, Y)$ is equivalent to maximizing $H(Y)$, since $H(Y|X)$ is noise.

The information maximization learning rule presented in [16] maximizes $H(Y)$ with respect to a weight matrix W, where $y_i = g(W^* x_i)$, and g is a logistic transfer function. The resulting learning rule shown in Eq. (4) 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:

$$\Delta W = \alpha[(W^{\mathrm{T}})^{-1} + y' x^{\mathrm{T}}), \tag{4}$$

where α is the learning rate.[1]

The way information maximization reduces statistical dependence can be understood 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 [22]:

$$H(Y) = H(y_1) + \ldots + H(y_n) - I(y_1, \ldots, y_n). \tag{5}$$

Inspection of equation Eq. (5) shows that maximizing $H(Y)$ encourages $I(y_1, \ldots, y_n)$ to be small. Thus, maximizing the joint entropy of the output encourages the individual outputs to move towards statistical independence. The mutual information is guaranteed to reach a minimum when the nonlinear transfer function g matches the cumulative distributions of the independent signals responsible for the data in X, up to a scaling and translation (Nadal and Parga, 1994). In most circumstances these distributions are unknown. However, for mixtures of super-Gaussian signals (meaning the probability distribution is more steeply peaked than a Gaussian), the logistic transfer function has been found sufficient to separate the signals into independent components [16]. As mentioned above, many natural signals including sound sources and

[1] Another version of this learning rule employs the *natural gradient*, which is the gradient multiplied by $W^{\mathrm{T}}W$, which regularizes the metric in the weight space [15]. Here the Hebbian learning term is between the input and the *natural gradient* of the output.

$$\Delta W = a(I + y' x^{\mathrm{T}} W^{\mathrm{T}})W.$$

measures of visual contrast have been shown to have a super-Gaussian distribution.

3.3. Sparse codes

Sharply peaked distributions are called sparse, since the vast majority of the responses are at or near zero, with rare responses at high values. There is a close relationship between ICA and sparse codes. Refer again to Eq. (5). If we hold the left-hand side constant, and then minimize $I(y_1,\ldots,y_n)$, this will also minimize $H(y_1)$ … $H(y_n)$. Thus minimizing the mutual information without loss of entropy in the joint distribution also minimizes the entropy of the marginal distributions. A minimum entropy distribution is a sharply peaked, sparse, distribution.

Barlow [7] advocated minimum entropy coding for this reason. When the *individual* neurons have minimum entropy, or sparse distributions, the redundancy between them is reduced. Olshausen and Field [55] presented a learning rule for image coding which explicitly implemented Barlow's concept of minimum entropy coding, in which the sparseness of the marginal distributions was maximized while minimizing the mean squared error of the reconstructed image. This learning rule can be related to Eq. (5), where we hold $H(Y)$ constant (no information loss), and this time minimize $H(y_1)$ … $H(y_n)$, which in turn minimizes $I(y_1,\ldots,y_2)$.

Indeed, it has been observed that when ICA is applied to natural signals, although the algorithm attempts to maximize the *joint* entropy, the marginal distributions become quite sparse [9,15]. Thus there is a difference in the shape of the optimal response distribution, depending on whether information transfer is maximized for an individual neural response or for a population code. For maximizing information transfer in an *individual* output (e.g. within the response of a single neuron) the optimal output distribution is a maximum entropy distribution, which in the case of limited dynamic range is a uniform distribution. In contrast, for maximizing information transfer in a *population code*, the response distributions of the individual neurons would tend to be sparse. Some researchers have reported a trend for neural responses to become more sparse higher in the visual system (e.g. [61,62,73]). Thus one might investigate whether neurons early in the visual system maximize information *within* a neural response, whereas neurons higher in the visual system maximize information in population codes.

3.4. Relationships of ICA to V1 receptive fields

A number of relationships have been shown between ICA and the response properties of visual cortical neurons. For example, Olshaussen and Field [54] showed that when the sparseness objective function described above was applied to natural images, the weights were local, spatially opponent edge filters similar to the receptive fields of V1 simple cells. In a related study, applying ICA to a set of natural images also produced V1-like receptive fields [15]. Conversely, Gabor filter responses to natural scenes have sparse distributions [27] and passing a set of natural scenes through a bank of Gabor filters followed by divisive normalization reduces dependencies [76]. When ICA is 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 dimensions [75]. Moreover, a two-layer ICA model learned translation invariant receptive fields related to complex cell responses [34].

4. Application to face recognition by computer

These learning principles have been applied to face images for face recognition [21,70]. Eigenfaces is essentially an unsupervised learning strategy that learns the second-order dependencies among the image pixels. It applies PCA to a set of face images. (See Fig. 5). Principal component solutions can be learned in neural networks with simple Hebbian learning rules (e.g. [53]). 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.

This approach, when it was first presented in the early 1990s performed considerably better than contemporary approaches that focused on measuring specific facial features and the geometric relationships between them. The eigenface approach lead to new research activity in the computer vision field focused on statistical learning in large populations of face images. The success of the eigenface

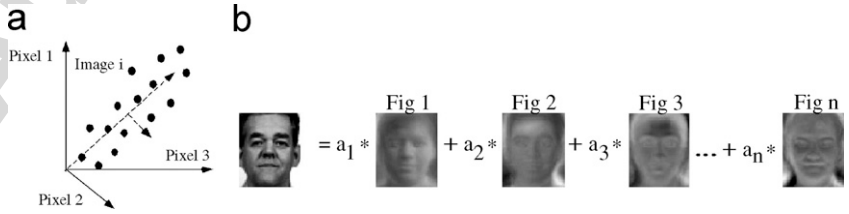

Fig. 5. Eigenfaces. (a) Each image is a point in a high-dimensional space defined by the grayvalue taken at each of the pixel locations. Principal component analysis rotates the axes to point in the directions of maximum covariance of the data. The axes are the eigenvectors of the pixelwise covariance matrix and are depicted by dashed lines. The faces are recoded by their coordinates with respect to the new axes. (b) The eigenvectors are themselves images, since they are vectors in pixel space, and form a set of basis images where each face is a weighted combination of the basis images. The weights a_i are the coordinates with respect to each eigenvector, and they form a representation, or feature vector, on which identity classification is performed.

approach may have been related to the fundamental coding strategy of learning the dependencies in a population of face images (Fig. 6).

PCA is an effective signal decomposition method which is widely used in part for its efficiency, as it does not require iterative optimization. Here we investigate face recognition performance when we take the information maximization concept a step further. PCA learns the second-order dependencies among the pixels but does not explicitly encode high-order dependencies. Second-order dependencies are pair-wise relationships between the pixels in the image database such as covariance, whereas high-order dependencies include functions of three or more pixels. Second-order dependencies are adequate to characterize Gaussian probability models, and PCA models maximize information transfer in the case where the input distributions are Gaussian. However, the preponderance of the statistical properties of natural images that have been measured are strikingly nonGaussian [24,27,33,41]. ICA models maximize information transfer under a much wider range of input distributions. PCA and ICA can be derived as generative models of the data, where PCA uses Gaussian sources, and ICA typically uses sparse sources. It has been shown that for many natural signals, ICA is a better model in that it assigns higher likelihood to the data than PCA [42].

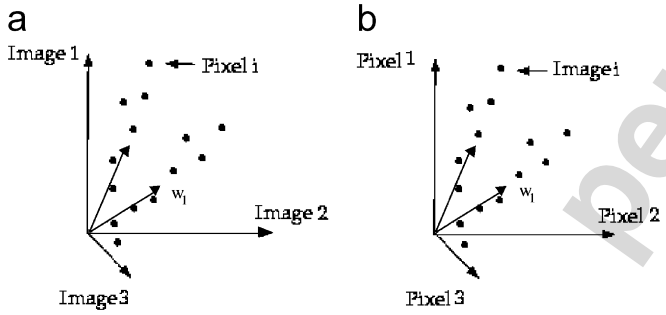

Fig. 6. Applying ICA to face images. (a) Architecture I. (b) Architecture II. Reprinted with permission from [9]. © 2002 IEEE.

Bartlett et al. [9] developed a representation for face recognition based on ICA. ICA learns the high-order dependencies in addition to the first- and second-order dependencies. This work compared ICA face representations to PCA to investigate whether explicitly encoding more of the dependencies would result in better recognition performance.

Bartlett, Movellan, and Sejnowski described two ways to apply ICA to face images. One approach is to treat each image as a variable and each pixel an observation. This approach, Architecture I, is illustrated in Fig. 7(a). Each pixel is plotted according to the grayvalue it assumes over a set of face images. This architecture has been employed in applications of ICA to analysis of fMRI [48]. In Architecture II, the images are treated as observations and the pixels as the variables (Fig. 7(b)). This architecture is more directly analogous to the PCA model illustrated in Fig. 5. Each image is a point in a high-dimensional space in which the dimensions are the pixels. The distinction between the two architectures is achieved by transposing the image matrix X during learning. The ICA algorithm treats the rows of X as a set of random variables, and the columns of X as the observations. Hence when the images are in the rows of X, the images are the variables and the pixels are the observations. When we transpose X, the pixels are treated as the random variables, and each face image is an observation.

The ICA model shown in Fig. 7 decomposes the images as $X = AS$ where A is an unknown mixing matrix and S is an unknown matrix of independent sources. The ICA learning rule (Eq. (4)) is applied to recover S by estimating $W = A^{-1}$. In both architectures, each face image is defined as a linear combination of a set of basis images. In Architecture I, S contains the basis images, and A contains

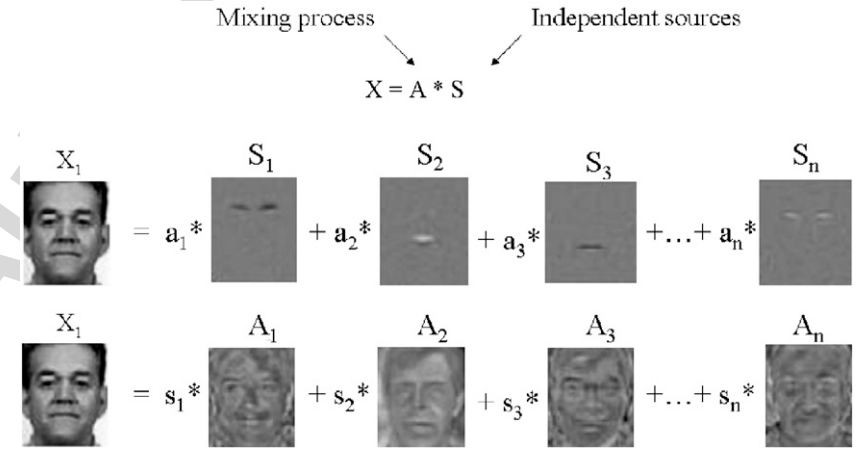

Fig. 7. Image synthesis model for the two architectures. The ICA model decomposes images as $X = AS$, where A is a mixing matrix and S is a matrix of independent sources. In Architecture I (top), the independent sources S are basis images and A contains the coefficients, whereas in Architecture II (bottom) the independent sources S are face codes (the coefficients) and A contains the basis images. Reprinted with permission from [8]. Copyright 2006, Elsevier.

the coefficients, whereas in Architecture II, S contains the coefficients and A contains the basis images. These two image synthesis models are a direct consequence of transposing the input matrix. In Architecture I, the face images are in the rows of X, the rows of S are in image space, and the coefficients for reconstructing each row of X from the rows in S are in the rows of the mixing-matrix A. In Architecture II, $X^{\mathrm{T}} = AS$. The face images are in the columns of X, the rows of A are in image space, and the coefficients for reconstructing each row of X from the rows in A are in the columns of S. Example basis images learned by the two architectures are shown in Fig. 7. The basis images learned in Architecture I are spatially local, whereas the basis images learned in Architecture II are global, or configural.

Bartlett, Movellan and Sejnowski 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 Fig. 8. Both ICA face representations outperformed PCA for the pictures taken on a different day. The robust recognition over time is particularly encouraging, since most applications of automated face recognition require recognizing face images collected at a different time point from the sample images, and is generally the more challenging condition for automated systems. ICA representations are in some ways better optimized for transmitting information in the presence of noise than PCA [8], and thus they may be more robust to variations such as lighting conditions, changes in hair, makeup, and facial expression which can be considered forms of noise with respect to the main source of information in our face database: the person's identity.

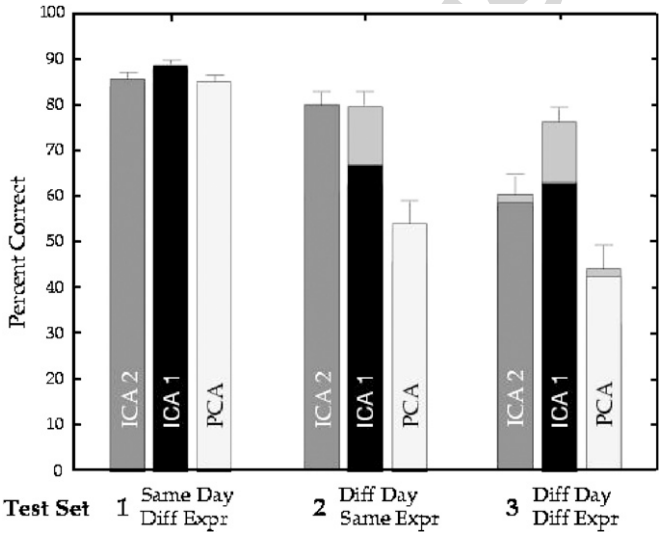

Fig. 8. Face recognition performance on the FERET database. ICA architctures I and II are compared to PCA (eigenfaces). Gray bars show improvement in performance following class-specific subselection of basis dimensions. Reprinted with permission from [9]. © 2002 IEEE.

When subsets of bases are selected, ICA and PCA define different subspaces. Bartlett, Movellan, and Sejnowski examined face recognition performance following subspace selection with both ICA and PCA. Dimensions 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 Fig. 8 show the improvement by subspace selection. ICA defined subspaces encoded more information about facial identity than PCA-defined subspaces. We also explored subselection 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. Here the focus was on representation. Both ICA and PCA as well as raw image pixels can provide the input to a variety of classifiers that employ supervised learning, including Fisher's linear discriminant [14], and support vector machines [59]. Indeed, class-specific projections of the ICA face codes using Fisher's linear discriminant was recently shown to be effective for face recognition [37].

4.1. Local representations versus factorial codes

Draper and colleagues [26] conducted a comparison of ICA and PCA on a substantially larger set of FERET face images consisting of 1196 individuals, and included a change in lighting condition which we had not previously tested. This study supported the finding that ICA outperformed PCA over time, as well as over changes in expression. ICA was also more robust to changes in lighting. ICA with architecture II obtained 51% accuracy

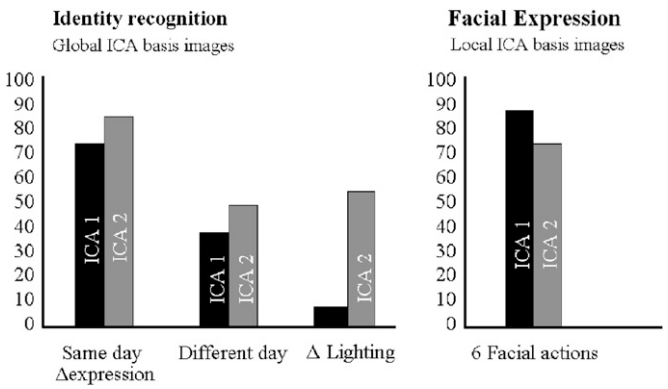

Fig. 9. Face recognition performance (percent correct) on a larger image set, based on the results of Draper and colleagues [26]. Figure reprinted with permission from [8]. Copyright 2006, Elsevier.

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 Fig. 9. According to arguments by Barlow [7] and Field [27] the sparse, factorial properties of the representation in Architecture II should be more optimal for face coding. Architecture II provides a factorial face code, in that each element of the feature vector is independent of the others (i.e. the coefficients are independent). The term factorial comes from the fact that when the marginal distributions are independent, the joint probability can be calculated as the product of the marginal probabilities. The representation in Architecture I is not factorial since it learns independent basis images, rather than independent coefficients [9]. 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. Architecture II had fewer training samples to estimate the same number of free parameters as Architecture I due to the difference in the way the input data was defined. With a substantially larger set of training images in the Draper study, the factorial representation of Architecture II emerged as more effective for identity recognition.

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 advantage for Architecture I only emerged following subspace selection using class variability ratios. The task was to recognize six 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. [32]).

Here we obtain better recognition performance when we define different basis sets for identity versus expression. In the two basis sets we switch what is treated as an observation versus what is treated as an independent variable for the purposes of information maximization.

5. Dependency learning and face perception

A number of perceptual studies support the relevance of dependency encoding to human face perception. A large body of work showed that unsupervised learning of second-order dependencies successfully models a number of aspects of human face perception including similarity, typicality, recognition accuracy, and other-race effects (e.g. [20,31,52]). 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 [30]. 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 [61,62].

Perceptual effects such as other-race effects are consistent with information maximization coding. For example, face discrimination is superior for same-race than other-race faces [77], which is consistent with a perceptual transfer function that is steeper for face properties in the high-density portion of the distribution in an individual's perceptual experience (i.e. same-race faces). See Fig. 10. Moreover, a morph stimulus that is half way on a physical continuum between a same-race face and an other-race face is typically perceived as more similar to the other-race face [36]. This is also consistent with the shape of the optimal transfer function, as illustrated in Fig. 10. A perceptual discrimination study by Parraga, Troscianko, and Tollhurst [58] supports information maximization in object perception. Sensitivity to small perturbations due to morphing was highest for pictures with natural second-order statistics, and degraded as the second-order statistics were made less natural. This is consistent with a transfer function with steeper slope where the probability density is highest in the environment.

Tanaka and colleagues [69] showed a related effect with typical and atypical faces. This study showed that morphs between typical and atypical parent faces appear to be more similar to the atypical parent. Fig. 10 shows an example of 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 (Fig. 11).

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. Bartlett and Tanaka [11] implemented an attractor network model of face perception, and indeed showed that atypical faces have larger basins of attraction. Inspection of this model provides further insights about the potential role of redundancy reduction in face perception. In this model, a decorrelation step was necessary in order to encode highly similar patterns in a Hopfield attractor network [35]. Decorrelation was necessary in order for each face to assume a distinct pattern of sustained activity. The atypicality bias in the model arose from this decorrelation process inherent to producing separate internal representations for correlated faces.

An alternative account of the atypicality bias is provided by information maximization and redundancy reduction. This account is compatible with the attractor network

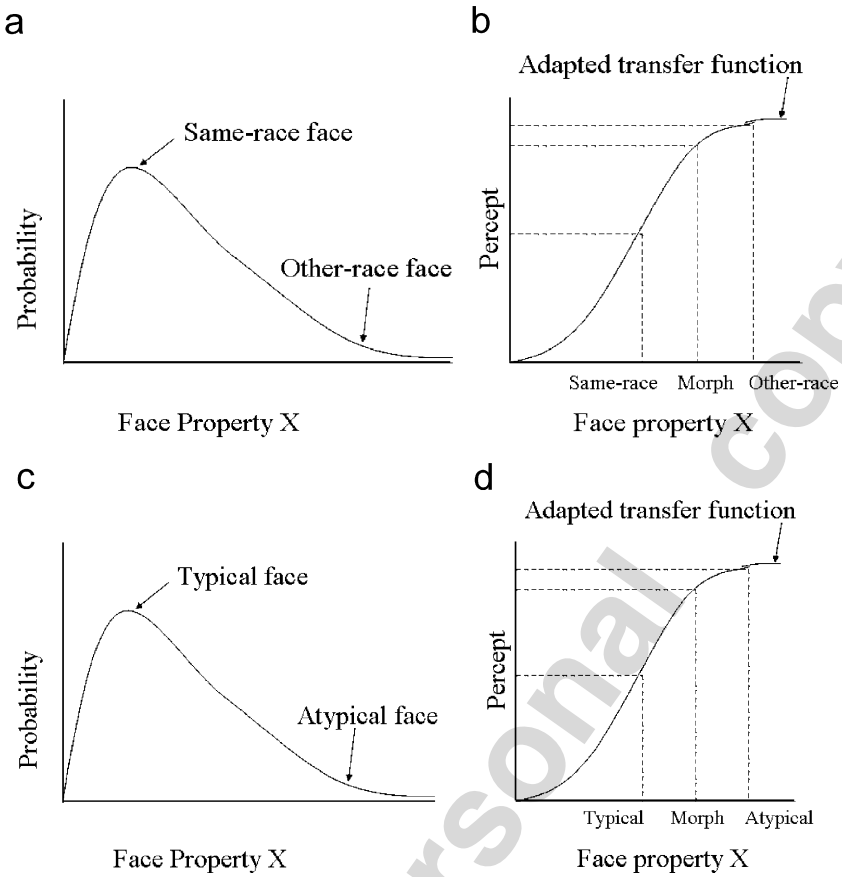

Fig. 10. Information maximization account of the other-race effect. (a) Example probability distribution for a face property such as eye shape, in which there is a higher probability density near the same-race than the other-race face. In cases of moderate exposure to other-race faces, probability distributions may be better characterized as multimodal. See Fig. 12 for a treatment of multimodal distributions. (b) Perceptual transfer function predicted by information maximization (the cumulative density of the distribution on the left). The slope of the transfer function is steeper near the same-race face, giving greater sensitivity. The percept of a physical 50% morph is mapped closer to the other-race than the same-race face because of the slope of the transfer function. (c–d) The same information maximization model accounts for the atypicality bias in the perception of morphed faces. (c) Example probability distribution for a face property for which there is a higher probability density near the typical face than the atypical face. (d) The percept of the 50% morph is mapped closer to the atypical face because of the shape of the transfer function.

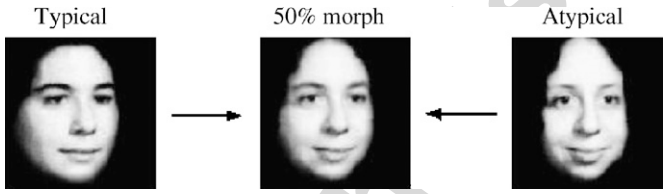

Fig. 11. Sample morph between typical and atypical parent faces, adapted from Tanaka [69].

hypothesis but more parsimonious. Fig. 10(c) shows the probability distribution of a physical face property X such as aspect ratio. A typical face has a value near the mean for this property, and an atypical face has a value that is in the tail. Fig. 10(d) shows the transfer function between the value of property X and the 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 small perturbations in the physical properties of atypical faces. This prediction was born out in a recent study [68]. Sensitivity to small perturbations in the face created by morphing was significantly lower for atypical than typical faces.

5.1. *Information maximization account of adaptation aftereffects*

Face adaptation studies (e.g. [36,40,51,81]) are consistent with information maximization on short time scales. For example, after adapting to a distorted face, a neutral face appears distorted in the opposite direction [81]. Similar effects have been shown for race, gender, and facial expressions [36]. Hence it appears that faces may be coded according to expected values along these dimensions. Adaptation to a nondistorted face does not make distorted

faces appear more distorted [81], which is consistent with an infomax account in which the estimated probability density is altered very little by exposure to a stimulus that already has high likelihood.

A number of researchers have proposed models of adaptation based on learning the statistics of recent visual inputs [6,76,79]. Adaptation aftereffects operate on a much shorter time scale than perceptual learning in the environment. The neural mechanisms may differ, but the information maximization principle may apply to both cases. Consider Fig. 12(a), in which the adapting face is a distorted face in the tail of the face distribution for face feature X such as aspect ratio. Adaptation could alter the estimated probability density of feature X in the immediate environment, as shown by the dashed curve. Fig. 12(b) shows the transfer functions predicted by information maximization on both the pre- and post-adaptation probability densities. 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. [81]). 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 of multimodal distributions such as Fig. 12(b), but that they do constrained optimization with a family of functions such as a sigmoid. This possibility is illustrated in Fig. 12(c). Here a sigmoid was fitted to the cumulative probability density of the adaptation distribution using logistic regression. Like the previous model, 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 Fig. 12(d). 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 the sensitivity curve towards the adapting stimulus.

Support for the information maximization model of adaptation comes from evidence that short-term adaptation reduces dependencies between neurons. This has been shown both physiologically [18] and using psychophysics [80]. Also, a V1 model that reduces dependencies between neurons by divisive normalization predicts numerous psychophysical and physiological adaptation effects [76]. Evidence for increases in sensitivity near the adapting stimulus has been mixed. In the case of light adaptation,

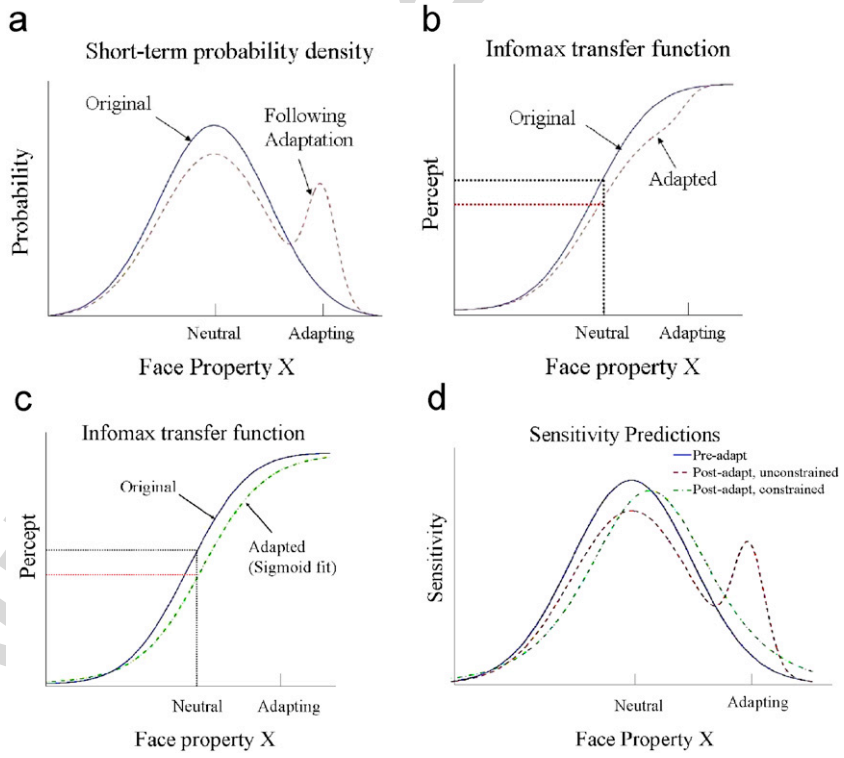

Fig. 12. (a) Example probability density of a physical face property such as aspect ratio before and after adaptation. Adaptation to a face with high aspect ratio changes the short-term probability density of aspect ratio. Extra density is added near the adapting stimulus, and density decreases elsewhere since it must sum to one. (b) Perceptual transfer function predicted by information maximization (the cumulative probability density). (c) Constrained optimization model in which the optimal transfer function following adaptation is approximated by a sigmoid. (d) Sensitivity functions predicted by the two transfer functions shown in b and c. 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 much smaller increase at the adapting stimulus, but an overall shift in the sensitivity curve towards the adapting stimulus.

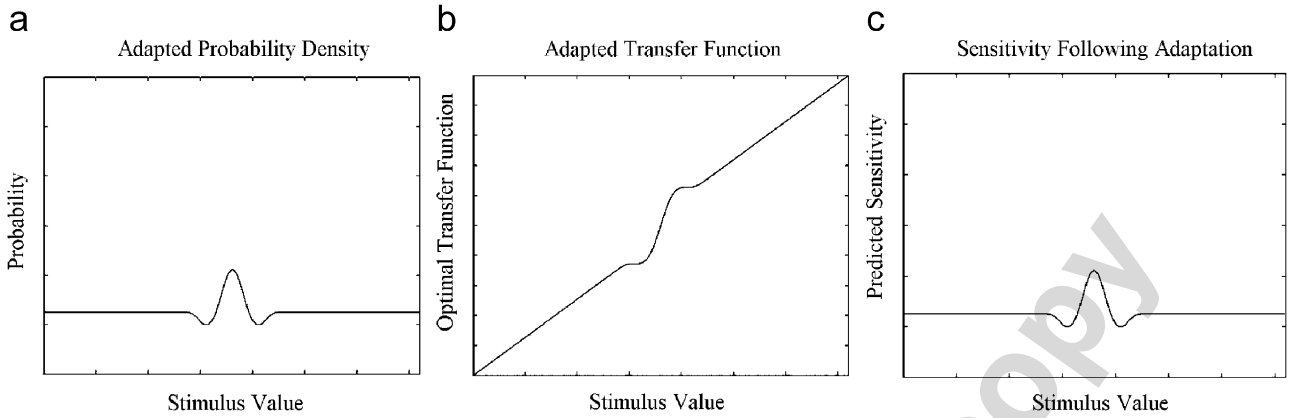

Fig. 13. Information maximization is consistent with the W-shaped sensitivity function such as observed psychophysically following motion adaptation. (a) Possible short-term probability density following adaptation, given a previously uniform distribution. Here we model the effect of adaptation to influence the probability density in a local range near the adapting stimulus value, as in [66]. The increase in probability near the adapting stimulus value is balanced by a decrease elsewhere within the local range, producing a Mexican-hat shape. (b) Optimal transfer function derived from the cumulative density of (a). (c) Sensitivity predictions derived from the first derivative of (b). Note that sensitivity increases near the adapting stimulus but also decreases for the stimuli immediately surrounding the adapting stimulus.

sensitivity clearly increases near the adapting light level [82]. Spectral sensitivity also increases near the adapting color [46,49]. There is also evidence that motion adaptation can improve both speed and direction discrimination [17,19,63]. However, whether similar improvements in pattern discrimination occur following adaptation to patterns remains controversial, and when these are reported, they seem somewhat weak [79]. The constrained optimization model of Fig. 12(c) may help reconcile these findings since it predicts a much smaller increase in sensitivity near the adapting stimulus. The model also predicts an overall shift in the sensitivity curve, which may be more easily detectable. Hence the model suggests that sensitivity should be measured at a range of stimulus values to look for such shifts.

Related accounts of adaptation have been discussed in terms of a generative model [25] and in a Bayesian framework [66]. As pointed out by [66], in cases where sensitivity has been shown to increase near the adapting stimulus, such as motion direction discrimination, sensitivity has also been shown to *decrease* in regions immediately surrounding the adapting stimulus, creating a W-shaped sensitivity curve (e.g. [19]). The information maximization model predicts such decreases in sensitivity to the surrounding stimuli. This can be seen in Fig. 12(d), but is more clear for approximately uniform distributions as illustrated in Fig. 13. Stocker and Simoncelli showed how this W-shaped sensitivity curve is consistent with a Bayesian model of adaptation which incorporates gain changes such as those described in this paper into the computation of the likelihood of the data. However a Bayesian model is not required to explain the W-shaped sensitivity curve, since it can also be explained directly by the information maximization model.

6. Discussion

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 [65] 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 [53]. Hence the Eigenface approach can be considered a form of Hebbian learning model, which performs information maximization under restricted conditions. In particular, PCA maximizes information transfer in the case where all of the signal distributions are Gaussian. ICA performs information maximization for a more general set of input distributions. The ICA learning algorithm employed here was developed from the principle of optimal information transfer in neurons with sigmoidal transfer functions. The learning rule contains a Hebbian learning term, but it is between the input and the *gradient* of the output. Section 4 showed that 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.

A number of perceptual studies support the relevance of dependency encoding to human face perception. Perceptual effects such as other-race effects are consistent with information maximization coding, where probability distributions are learned over long-term perceptual experience. Face adaptation studies (e.g. [36,40,51,81]) are consistent with information maximization on short time scales. Unsupervised learning of second-order dependencies (PCA) successfully models a number of aspects of human face perception including similarity, typicality, recognition accuracy, and other-race effects (e.g. [20,31,52]). 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 [30]. 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 [61,62].

Two models of information maximization in adaptation were presented, one in which the visual system learns a high-order transfer functions to match cumulative probability densities of multimodal distributions, and another in which the cumulative probability density is approximated with a family of functions such as a sigmoid. The second model does not predict large increases in sensitivity near the adapting stimulus, and may help account for the weak evidence for such increases following adaptation. This second model suggests that sensitivity should be measured at a range of stimulus values to look for shifts in the full sensitivity curve.

Desirable filters may be those that are adapted to the patterns of interest and capture interesting structure [43]. The more the dependencies that are encoded, the more structure that is learned. Information theory provides a means for capturing interesting structure. Information maximization leads to an efficient code of the environment, resulting in more learned structure. Such mechanisms predict neural codes in both vision [15,54,75] and audition [42]. The research presented here found that face representations in which high-order dependencies are separated into individual coefficients gave superior recognition performance to representations which only separate second-order redundancies.

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. [10,12,28,56,67,78]). Support for learning spatiotemporal dependencies in the visual system comes from observations that optimal neural filters match spatiotemporal contrast sensitivity functions [71], as well as responses in the LGN to natural movies [23]. Moreover, spatio-temporal ICA on movies of natural images produces components that resemble the direction-selective receptive fields of V1 neurons [72]. There is also evidence that receptive fields in the primate anterior temporal lobe are plastically modified by new temporal dependency structure in visual experience. After macaques were presented a sequence of fractal patterns for 6 weeks, the responses of AIT cells to neighboring stimuli in the sequence became correlated, and the correlation reduced as the distance in the sequence increased [50].

The information maximization algorithm employed in this work assumed that the pixel values in face images were generated from a linear mixing process. This linear approximation has been shown to hold true for the effect of lighting on face images [29]. Other influences, such as changes in pose and expression may be linearly approximated only to a limited extent. Nevertheless, filters resembling simple cells in the primary visual cortex were learned by linear models [15,54]. Nonlinear ICA in the absence of prior constraints is an ill-conditioned problem, but some progress has been made by assuming a linear mixing process followed by parametric nonlinear functions [39,83]. An algorithm for nonlinear ICA based on kernel methods has also recently been presented [4]. Kernel methods have already shown to improve face recognition performance with PCA and Fisherfaces [84], and promising results have recently been presented for face recognition with kernel-ICA [45]. Another promising approach to nonlinear ICA is through a model of divisive normalization in V1 [76]. In this approach, weights on neighboring filters in space, scale, and orientation are learned using an objective function that is related to maximizing the sparseness of each normalized response distribution across a set of natural images. Future work includes exploring such divisive normalization models for representing faces.

The information maximization models discussed here do not address the case of noise. When the slope of the transfer function is increased, errors are magnified proportionally. Von der Twer and Macleod [74] argued that the optimal transfer function should simultaneously maximize information transfer while minimizing a measure of error. They showed that shallower transfer functions than the ones learned by information maximization, proportional to the cube root of the cumulative pdf, optimize information transmission in the presence of noise. They showed that this model accounts well for the spectral sensitivities of the primate color opponent system. Moreover, psychophysical estimates of the slope of the transfer function matched roughly to the cube root of the probability density of the environmental input values along the red–green, blue–yellow and light-dark dimensions of color space [47]. The infomax transfer function was shown to be a special case of von der Twer and Macleod's general objective function in which the error function was a constant. In other words, the infomax objective function

amounts to minimizing errors without regard to their size. Hence another avenue of research is to explore face representations using optimization functions such as in Ref. [74] which take the size of the error into account.

References

[1] J.J. Atick, Could information theory provide an ecological theory of sensory processing?, Network 3 (1992) 213–251.

[2] J.J. Atick, A.N. Redlich, What does the retina know about natural scenes?, Neural Comput. 4 (1992) 196–210.

[3] F. Attneave, Some informational aspects of visual perception, Psychol. Rev. 61 (1954) 183–193.

[4] F.R. Bach, M.I. Jordan, Kernel independent component analysis, in: Proceedings of the Third International Conference on Independent Component Analysis and Signal Separation, 2001.

[5] H. Barlow, What is the computational goal of the neocortex?, in: C. Koch (Ed.), Large Scale Neuronal Theories of the Brain, MIT Press, Cambridge, MA, 1994, pp. 1–22.

[6] H. Barlow, P. Foldiak, Adaptation and decorrelation in the cortex, in: C.M.R. Durbin, G. Mitchinson (Eds.), The Computing Neuron, Addison-Wesley, Reading, MA, 1990, pp. 54–72.

[7] H.B. Barlow, Unsupervised learning, Neural Comput. 1 (1989) 295–311.

[8] M.S. Bartlett, J.R. Movellan, T.J. Sejnowski, Face modeling by information maximization, in: R.C.a.Y. Zhao (Ed.), Face Processing: Advanced Modeling and Methods, Elsevier, Amsterdam, 2006, pp. 219–253.

[9] M.S. Bartlett, J.R. Movellan, T.J. Sejnowski, Face recognition by independent component analysis, IEEE Trans. Neural Networks 13 (6) (2002) 1450–1464.

[10] M.S. Bartlett, T.J. Sejnowski, Learning viewpoint invariant face representations from visual experience in an attractor network, Network 9 (3) (1998) 399–417.

[11] M.S. Bartlett, J.W. Tanaka, An attractor field model of face representations: effects of typicality and image morphing, in: Proceedings of the Psychonomics Society Satellite Symposium on Object Perception and Memory (OPAM), 1998.

[12] S. Becker, Implicit learning in 3D object recognition: the importance of temporal context, Neural Comput. 11 (2) (1999) 347–374.

[13] S. Becker, M. Plumbley, Unsupervised neural network learning procedures for feature extraction and classification, J. Appl. Intell. 6 (1996) 1–21.

[14] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.

[15] A.J. Bell, T.J. Sejnowski, The independent components of natural scenes are edge filters, Vision Res. 37 (23) (1997) 3327–3338.

[16] A.J. Bell, T.J. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, Neural Comput. 7 (6) (1995) 1129–1159.

[17] P.J. Bex, S. Bedingham, S.T. Hammett, Apparent speed and speed sensitivity during adaptation to motion, J. Opt. Soc. Am. A 16 (12) (1999) 2817–2824.

[18] M. Carandini, J.A. Movshon, D. Ferster, Pattern adaptation and cross-orientation interactions in the primary visual cortex, Neuropharmacology 37 (4–5) (1998) 501–511.

[19] C.W. Clifford, P. Wenderoth, Adaptation to temporal modulation can enhance differential speed sensitivity, Vision Res. 39 (26) (1999) 4324–4332.

[20] G. Cottrell, et al., Is all face processing holistic? The view from UCSD, in: M.W.a.J. Townsend (Ed.), Computational, Geometric, and Process Perspectives on Facial Cognition: Contexts and Challenges, Erlbaum, London, 2000.

[21] G. Cottrell, J. Metcalfe, Face, gender and emotion recognition using Holons, in: D. Touretzky (Ed.), Advances in Neural Information Processing Systems, Morgan Kaufmann, San Mateo, CA, 1991, pp. 564–571.

[22] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley, New York, 1991.

[23] Y. Dan, J.J. Atick, R.C. Ried, Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory, J. Neurosci. 16 (1996) 3351–3362.

[24] J.G. Daugman, Entropy reduction and decorrelation in visual coding by oriented neural receptive fields, IEEE Trans. Biomed. Eng. 36 (1) (1989) 107–114.

[25] P. Dayan, M. Sahani, G. Deback, Adaptation and unsupervised learning, in: S.T.S. Becker, K. Obermayer (Eds.), Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 2002.

[26] B.A. Draper, et al., Recognizing faces with {PCA} and {ICA}. Computer vision and image understanding,, Face Recognition 91 (2003) 115–137 (special issue).

[27] D.J. Field, What is the goal of sensory coding?, Neural Comput. 6 (1994) 559–601.

[28] P. Foldiak, Learning invariance from transformation sequences, Neural Comput. 3 (1991) 194–200.

[29] P. Hallinan, A deformable model for face Recognition under arbitrary lighting conditions, Ph.D. Thesis, Harvard University, 1995.

[30] P. Hancock, Alternative representations for faces, in: Proceedings of the British Psychological Society, Cognitive Section, University of Essex, September 6–8, 2000.

[31] P.J.B. Hancock, A.M. Burton, V. Bruce, Face processing: human perception and principal components analysis, Memory Cognition 24 (1996) 26–40.

[32] J. Haxby, E. Hoffman, M. Gobbini, Trends Cognitive Sci. 4 (2000) 223–233.

[33] J. Huang, D. Mumford, Statistics of natural images and models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1999, pp. 541–547.

[34] A. Hyvarinen, P.O. Hoyer, A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images, Vision Res. 41 (18) (2001) 2413–2423.

[35] I. Kanter, H. Sompolinsky, Associative recall of memory without errors, Phys. Rev. A 35 (1987) 380–392.

[36] D. Kaping, P. Duhamel, M. Webster, Adaptation to natural face categories, J. Vision 2 (10) (2002) 128.

[37] J. Kim, J. Choi, J. Yi, Face recognition based on ICA combined with FLD, in: Proceedings of the European Conference on Computer Vision, 2002, pp. 10–18.

[38] S. Laughlin, A simple coding procedure enhances a neuron's information capacity, Z. Nat. 36 (1981) 910–912.

[39] T.-W. Lee, B.U. Koehler, R. Orglmeister, Blind source separation of nonlinear mixing models, in: Proceedings of the IEEE International Workshop on Neural Networks for Signal Processing, Florida, 1997, pp. 406–415.

[40] D. Leopold, et al., Prorotype-referenced shape encoding revealed by high-level aftereffects, Nat. Neurosci. 4 (2001) 89–94.

[41] M. Lewicki, Efficient coding of natural sounds, Nat. Neurosci. 5 (2002) 356–363.

[42] M. Lewicki, B. Olshausen, Probabilistic framework for the adaptation and comparison of image codes, J. Opt. Soc. Am. A 16 (7) (1999) 1587–1601.

[43] M. Lewicki, T.J. Sejnowski, Learning overcomplete representations, Neural Comput. 12 (2) (2000) 337–365.

[44] R. Linsker, Self-organization in a perceptual network, Computer 21 (3) (1988) 105–117.

[45] Q. Liu, et al., Modeling face appearance with nonlinear independent component analysis, in: Proceeding of the Sixth International Conference on Automatic Face and Gesture Recognition, 2004.

[46] J.M. Loomis, T. Berger, Effects of chromatic adaptation on color discrimination and color appearance, Vision Res. 19 (8) (1979) 891–901.

[47] D.I.A. MacLeod, Colour discrimination, colour constancy and natural scene statistics, in: J.D. Mollon, J. Pokorny, K. Knoblauch (Eds.), Normal and Defective Colour Vision, Oxford University Press, New York, 2003 (The Verriest lecture).
[48] M.J. McKeown, et al., Analysis of f{MRI} by decomposition into independent spatial components, Hum. Brain Mapp. 6 (3) (1998) 160–188.
[49] E. Miyahara, V.C. Smith, J. Pokorny, How surrounds affect chromaticity discrimination, J. Opt. Soc. Am. A 10 (4) (1993) 545–553.
[50] Y. Miyashita, Neuronal correlate of visual associative long-term memory in the primate temporal cortex, Nature 335 (27) (1988) 817–820;
J.-P. Nadal, N. Parga, Non-linear neurons in the low-noise limit: a factorial code maximises information transfer, Network 4 (1994) 295–312.
[51] M. Ng, et al., Selective tuning of face perception, J. Vision 3 (9) (2003) 106a (Abstract).
[52] A. O'Toole, et al., Structural aspects of face recognition and the other race effect, Memory Cognition 22 (2) (1994) 208–224.
[53] E. Oja, Neural networks, principal components, and subspaces, Int. J. Neural Syst. 1 (1989) 61–68.
[54] B.A. Olshausen, D.J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images, Nature 381 (1996) 607–609.
[55] B.A. Olshausen, D.J. Field, Natural image statistics and efficient coding, Network: Comput. Neural Syst. 7 (2) (1996) 333–340.
[56] R. O'Reilly, M. Johnson, Object recognition and sensitive periods: a computational analysis of visual imprinting, Neural Comput. 6 (1994) 357–389.
[57] A. Papoulis, Probability, Random Variables, and Stochastic Processes, McGraw-Hill, New York, 1991.
[58] C.A. Párraga, T. Troscianko, D.J. Tollhurst, The human visual system is optimised for processing the spatial information in natural visual images, Curr. Biol. 10 (1) (2000) 35–38.
[59] J. Phillips, Support vector machines applied to face recognition, in: T.L.S. Sola, K. Muller (Eds.), Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 1999, pp. 803–809.
[60] J. Rissanen, Modeling by shortest data description, Automatica 14 (1978) 465–471.
[61] E.T. Rolls, N.C. Aggelopoulos, L. Franco, A. Treves, Information encoding in the inferior temporal cortex: contributions of the firing rates and correlations between the firing of neurons, Biol. Cybern. 90 (2004) 19–32.
[62] E.T. Rolls, M.J. Tovee, Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex, J. Neurophysiol. 73 (2) (1995) 713–726.
[63] P.R. Schrater, E.P. Simoncelli, Local velocity representation: evidence from motion adaptation, Vision Res. 38 (24) (1998) 3899–3912.
[64] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423 623–656.
[65] E.P. Simoncelli, B.A. Olshausen, Natural image statistics and neural representation, Annu. Rev. Neurosci. 24 (2001) 1193–1216.
[66] A.A. Stocker, E.P. Simoncelli, Sensory adaptation within a Bayesian framework, in: Y. Weiss, B. Scholkopf, B. Platt (Eds.), Advances in Neural Information Processing Systems, vol. 18, MIT Press, Cambridge, MA, 2006, pp. 1291–1298.
[67] J.V. Stone, Learning perceptually salient visual parameters using spatiotemporal smoothness constraints, Neural Comput. 8 (7) (1996) 1463–1492.
[68] J.W. Tanaka, O. Corneille, Atypicality bias in face and object perception. Further tests of an attractor model, Perception & Psychophysics, in press.
[69] J.W. Tanaka, et al., Mapping attractor fields in face space: the atypicality bias in face recognition, Cognition 68 (3) (1998) 199–220.
[70] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cognitive Neurosci. 3 (1) (1991) 71–86.
[71] J.H. van Hateren, Spatiotemporal contrast sensitivity of early vision, Vision Res. 33 (1993) 257–267.
[72] J.H. van Hateren, D.L. Ruderman, Independent component analysis of natural image sequences yields spatiotemporal filters similar to simple cells in primary visual cortex, Proc. R. Soc. London Ser. B 265 (1412) (1998) 2315.
[73] W.E. Vinje, J.L. Gallant, Sparse coding and decorrelation in primary visual cortex during natural vision, Science 287 (5456) (2000) 1273–1276.
[74] T. von der Twer, D.I. MacLeod, Optimal nonlinear codes for the perception of natural colours, Network 12 (3) (2001) 395–407.
[75] T. Wachtler, T.-W. Lee, T.J. Sejnowski, The chromatic structure of natural scenes, J. Opt. Soc. Am. A 18 (1) (2001) 65–77.
[76] M.J. Wainwright, O. Schwartz, E.P. Simoncelli, Natural image statistics and divisive normalization: modeling nonlinearities and adaptation in cortical neurons, in: B.O.R. Rao, M. Lewicki (Eds.), Statistical Theories of the Brain, 2002.
[77] T.M. Walker, J.W. Tanaka, An encoding advantage for own-race versus other-race faces, Perception 23 (9) (2003) 1117–1125.
[78] G. Wallis, E.T. Rolls, Invariant face and object recognition in the visual system, Prog. Neurobiol. (Oxford) 51 (2) (1997) 167–194.
[79] M. Webster, J. Werner, D. Field, Adaptation and the phenomenology of perception, in: C.C.G. Rhodes (Ed.), Fitting the Mind to the World: Adaptation and Aftereffects in High Level Vision, Oxford University Press, Oxford, 2005.
[80] M.A. Webster, Human color vision and its adaptation, Network: Comput. Neural Syst. 7 (1996) 587–634.
[81] M.A. Webster, O.H. Maclin, Figural aftereffects in the perception of faces, Psychon. Bull. Rev. 6 (4) (1999) 647–653.
[82] P. Whittle, Brightness, discriminability, and the 'Crispening Effect', Vision Res. 32 (1992) 1493–1507.
[83] H.H. Yang, S.-I. Amari, A. Cichocki, Information-theoretic approach to blind separation of sources in non-linear mixture, Signal Process. 64 (3) (1998) 291–3000.
[84] M. Yang, Face recognition using kernel methods, in: T.D.a.B.S.a.Z. Ghahramani (Ed.), Advances in Neural Information Processing Systems, 2002.

Dr. Bartlett is Assistant Research Professor at the Institute for Neural Computation, UCSD, where she co-directs the Machine Perception Lab. She studies learning in vision, with application to face recognition and expression analysis. She has authored over 30 articles in scientific journals and refereed conference proceedings, as well as a book, *Face Image Analysis by Unsupervised Learning*, published by Kluwer in 2001. Dr. Bartlett obtained her Bachelor's degree in Mathematics in 1988 from Middlebury College, and her Ph.D. in Cognitive Science and Psychology from University of California, San Diego, in 1998. Her thesis work was conducted with Terry Sejnowski at the Salk Institute. She has also published papers in visual psychophysics with Jeremy Wolfe, neuropsychology with Jordan Grafman, perceptual plasticity with V.S. Ramachandran, machine learning with Javier Movellan, automatic recognition of facial expression with Paul Ekman, cognitive models of face perception with Jim Tanaka, and the visuo-spatial properties of faces and American Sign Language with Karen Dobkins. She has given numerous invited symposia to international audiences in both machine vision and cognitive neuroscience. Dr. Bartlett has been highly active in organizing workshops and conferences in the areas of learning in vision and affective computing. She co-organized two major conferences, the 11th European Conference on Visual Perception, Bristol, England, and the Third International Conference on Development and Learning in San Diego, California, as well as numerous workshops at conferences such as Advances in Neural Information Processing Systems, and International Conference on Computer Vision.