

Images, Frames, and Connectionist Hierarchies

Peter Dayan

dayan@gatsby.ucl.ac.uk

*Gatsby Computational Neuroscience Unit, University College London,
London WC1N@3AR*

The representation of hierarchically structured knowledge in systems using distributed patterns of activity is an abiding concern for the connectionist solution of cognitively rich problems. Here, we use statistical unsupervised learning to consider semantic aspects of structured knowledge representation. We meld unsupervised learning notions formulated for multilinear models with tensor product ideas for representing rich information. We apply the model to images of faces.

1 Introduction ---

What do we know when we know the story of *Moby Dick* or the face of an acquaintance? This question, or, more formally, that of competently representing objects with hierarchical structure in connectionist systems, has a critical part to play in addressing a range of pressing challenges in computational cognitive science. Various ingenious suggestions have been made (many starting and building from the foundation provided by the seminal collection of papers in Hinton, 1991), involving a wide range of computationally sophisticated mechanisms.

Much of this work is strongly motivated by aspects of predicate logic. Thus, it is concerned with discrete literals and logical terms or propositions linking them, and specifically with the idea that it should be possible to fashion a representation freshly on the fly for essentially arbitrary concepts. This rather episodic view accurately characterizes some cognitive tasks, particularly those associated with linguistics. However, there are many other tasks for which there is substantial structured semantic knowledge, that is, (typically hierarchically) structured networks of statistical relationships among a set of entities. This semantic structure provides a framework within which episodic information should be viewed. The example we use in this article is visual images of faces. There are numerous strong statistical constraints in such images—for one simple instance, the close similarity of the two eyes or two ears—and it is these that we seek to capture.

1.1 Unsupervised Learning. The requirements for this view amount to finding this semantic structure and using it to practical effect. We seek

to do both of these in a connectionist context, with distributed representations and without explicit pointers and the like. One obvious direction to turn for ideas and methods is statistical unsupervised learning algorithms (see Hinton, 1990; Rao, Olshausen, & Lewicki, 2002), which are explicitly designed to extract and represent semantic structure of various sorts, and whose connectionist credentials are burnished by their widespread use for modeling the nature and development of the tuning properties of cortical neurons. However, bar a few exceptions (notably for our work, Tenenbaum & Freeman, 2000, and, under a somewhat whiggish interpretation, Grimes & Rao, 2005), they have not been much applied to hierarchical structure.

Versions of unsupervised learning based on density estimation can be viewed in the informal terms of characterizing the statistical structure of the input patterns in terms of low-dimensional manifolds and finding a coordinate system that parameterizes these manifolds. For instance, G. Hinton (personal communication, 1994) has estimated that images of faces live in a roughly 30- to 40-dimensional implicit space, embedded in the huge numbers of dimensions of pixel-based inputs. Edelman (1999) presents an excellent discussion of this sort of representation, albeit somewhat divorced from the context of statistical unsupervised learning.

The way the manifolds are embedded in the space in which the input lives, captures the overall statistical constraints among the collection of patterns. The manifolds are useful in that individual examples can be represented in terms of their coordinates. The resulting representation system, if learned correctly, provides an optimally compact representation for new inputs drawn from the same distribution. It explicitly does not, however, provide sensible coordinates for inputs that come from different distributions. It is intended to solve a different problem from that of representing arbitrary episodic structure. Unsupervised learning algorithms that employ strong priors can make strong claims for the manifolds and coordinate systems they extract, in the sense of finding things like underlying independent structure in the collection of examples.

In sum, we consider the problem of discovering and using semantic structure in domains in which it has inherently hierarchical forms. Although it is an important task for the future, in this letter, we do not seek to find the hierarchy itself (for images, this is provided naturally by the focus of attention) but rather to elucidate its representational implications.

1.2 Representations for Visual Objects. Connectionist representations sit at levels of detail and abstraction above those of neurally realizable codes. We focus on a relatively narrow and concrete question about the representation of hierarchical structure in a particular domain, and therefore adopt three broad constraints and (gross) simplifications associated with the sort of visual images we use: segmentation, invariance, and distributed representations. Further, again to focus on representation, we do not attempt to solve the challenging general problem of detection and classification of

faces and other objects in images, a task that over the past several years has attracted a number of powerful and probabilistic approaches (Burl, Leung, & Perona, 1995; Burl, Weber, & Perona, 1998; Schiele & Crowley, 1996, 2000; Fei-Fei, Fergus, & Perona, 2003; Fergus, Perona, & Zisserman, 2003; Liebe & Schiele, 2003, 2004; Schneiderman & Kanade, 2004; Amit & Trouvé, 2005; Sudderth, Torralba, Freeman, & Willsky, 2005; Crandall, Felzenszwalb, & Huttenlocher, 2005). We discuss the relationship between our work and these ideas later.

Issues of segmentation and invariance mostly have to do with pre-processing. We help ourselves to a mechanism capable of extracting the elements of a scene (e.g., a whole face, an eye, a nose) at appropriate scales, in normalized coordinates. This is exactly the intent of Olshausen, Anderson, & Van Essen, (1993) explicit shifter circuit, and the recent architecture of Amit and Mascaró (2003), which powerfully integrates detection and recognition. It also underlies von der Malsburg's (1988) dynamic link architecture, and indeed has some resonances in the more bottom-up invariance sought in architectures such as the MAX model (Riesenhuber & Poggio, 1999). Even in the face of the limited evidence about basis functions associated with the focus of attention (Connor, Gallant, Preddie, & Van Essen, 1996) for achieving an equivalent (Pouget & Sejnowski, 1997) of shifting, this is obviously a large simplification. We justify it on the basis of our key interest in the question of representation.

We also make the great simplification of restricting the manifold to be a mixture of factor analysers (Hinton, Dayan, & Revow, 1997). This does lead to a distributed code, but one that is obviously far too simple to reflect faithfully the sort of population code representation that we might legitimately expect in the brain (Pouget, Dayan, & Zemel, 2000).

1.3 Tensors and Distributed Representations. Smolensky (1990) suggested that tensor products are the natural means for representing and manipulating structured knowledge that is represented as distributed patterns of activity over multiple units. This idea has exerted significant influence over a wealth of subsequent work in the field, including, for instance, the approaches of Plate (1995, 2003) and Gayler (1998), who have studied generalizations and simplifications of tensor products, and also the community working on recursive autoassociative memories (Hinton, 1990; Pollack, 1990; Sperduti, 1994).

Our work, which is an extension of Riesenhuber and Dayan (1996), also fits comfortably into this tradition, albeit in the context of semantic statistical ideas of unsupervised learning and thus the multilinear modeling framework of Tenenbaum and Freeman (2000; see also Vasilescu & Terzopoulos, 2002, 2003, and Grimes & Rao, 2005). Compared with Riesenhuber and Dayan (1996), we consider a much richer domain of visual objects (Banz & Vetter, 1999), and employ a more powerful unsupervised learning algorithm that can also automatically cluster the objects into classes.

In section 2 we describe the statistics of a structured domain, using a form of discrete, multiscale representation of images of faces as an example. We also describe the multilinear, unsupervised learning model that we employ to capture these statistics. In section 3, we generalize this approach to encompass unsupervised clustering of separate object classes or subclasses. Finally, in section 4, we consider how our model fits in with other ideas on structured knowledge representation and present some more speculative notions about domains rather far removed from face images.

2 Multilinear Models

The critical representational notion in this article is that hierarchically structured image objects should be considered as mappings from some form of generalized focus of attention or eye position \mathbf{e} to a form of observation \mathbf{x} that would be made at that focus of attention, that is:

$$\begin{array}{l} \text{image object : attentional position} \Rightarrow \text{observation.} \\ \mathcal{I} : \qquad \qquad \qquad \mathbf{e} \Rightarrow \mathbf{x} \end{array} \quad (2.1)$$

Connor et al.'s (1996) findings on the effects of the focus on attention on receptive fields in area V4 in visual cortex underlay Riesenhuber and Dayan's (1996) suggestion of a model of exactly this form (see also Salinas & Abbott, 1997). However, mappings of this sort date back at least to ideas on the interactions between action and observation (see, e.g., the extensive discussion in Bridgeman, van der Heijden, & Velichkovsky, 1994). In terms of a statistical generative model (MacKay, 1956; Neisser, 1967; Grenander, 1976–1981; Mumford, 1994; Hinton & Zemel, 1994; Dayan, Hinton, Neal, & Zemel, 1995; Olshausen & Field, 1996; Hinton & Ghahramani, 1997), often ascribed to feedback connections between cortical areas, the mapping in equation 2.1 suggests that correlations among the observations \mathbf{x} should be explained by two structural features of the inputs: the existence of multiple attentional foci for the same underlying object and the semantic (and episodic) structure of the image objects \mathcal{I} .

Figure 1 illustrates one way to conceive of the generative structure in the images of faces and also shows how we generated the training data for the letter. We used the face images from Blanz and Vetter (1999) together with fiducial markers (T. Vetter, personal communication, May 2005; M. Riesenhuber, personal communication, May 2005; Riesenhuber, Jarudi, Gilad, & Sinha, 2004) that locate particular features (such as the pupils of the eyes) in each image. The markers for the faces in the database were labeled by hand. As discussed in section 1, doing this automatically is an important task for preprocessing and is one of the key computations in models such as Olshausen et al.'s (1993) and Amit and Mascaró's (2003) shifter models; however, we do not model it explicitly here.

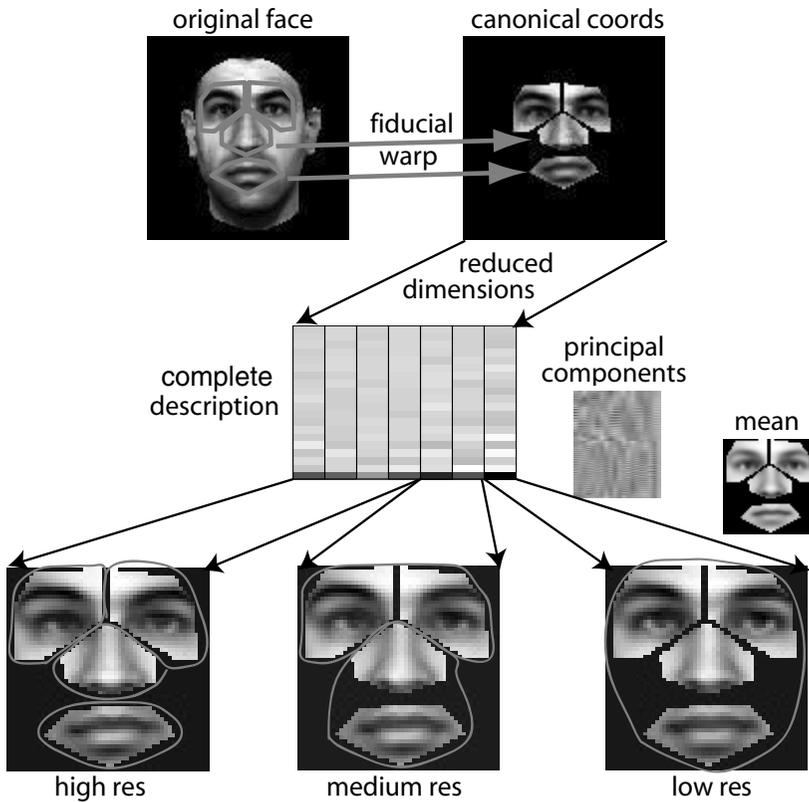


Figure 1: Hierarchical image decomposition. The left and right eyes, the nose, and the mouth of 190 faces from the Blanz and Vetter (1999) database (an example is shown in the top left, with the segments defined by fiducial markers) are warped into the common reference frame specified by one of the faces. Seven different images at three different resolutions are defined for each warped face (the four separate subparts at the highest resolution, the two eyes together, and the nose and mouth together at medium resolution, and all of them combined at the lowest resolution), and are projected onto the top 20 principal components of the seven separate covariance matrices. The 1488 pixels are therefore represented three times over (once per resolution) in 140 numbers (middle rectangular block). The subparts of the face and the face itself can be reconstructed from these coefficients to quite high fidelity (lower figures; with the irregular outlines showing which parts defined the foci of attention).

The markers create a linear object class representation of the faces (Vetter & Poggio, 1997; Beymer & Poggio, 1996), which allows them to be warped into a common reference frame, which we arbitrarily define based on the first face in the database (we could equally well have used the average face).

For simplicity, we concentrate in this article on the two eyes, the nose, and the mouth.

The top left-hand image in Figure 1 is an example face from the database. The irregular lines delimit regions containing the eyes, nose, and mouth, using the fiducial markers. These regions are then separately warped into canonical coordinates, defined as those of the eyes, nose, and mouth of the first face in the database. The image at the top right of the figure shows the result of this warping. The full images in the database are defined over $100 \times 100 = 10,000$ pixels; in the canonical representation, right eye, left eye, nose, and mouth are defined by 433, 394, 310, and 351 pixels, respectively (since these are the sizes of these features in the first face).

We assume that subjects can determine their focus of attention at one of three resolutions and thereby to seven discrete parts or subparts. At the highest resolution, the four individual elements of the face can be separately attended; at a medium resolution, either the two eyes or the nose and mouth together can be selected; at the lowest resolution, all four parts are represented collectively. The difference in resolution arises since items in the focus of attention are represented in a fixed size structure, so, for instance, the fidelity with which the full face can be represented is roughly a quarter that of the individual elements. In practice, we create this fixed structure by projecting the full input onto a fixed number d ($d = 20$ in the figure) of the principal eigenvectors of their covariance matrices (using separate covariance matrices for each of the seven resolutions). In terms of the relationship in equation 2.1, an observation x is the reduced, d -dimensional description of one element of a face at one resolution.

Principal component analysis (PCA) is exactly the outcome of the simplest Hebbian unsupervised learning algorithm applied to the warped images (Linsker, 1988). Performing PCA is sensible because of the linear class structure created by the fiducial markers (Vetter & Poggio, 1997; Beymer & Poggio, 1996). In our highly simplified description, we consider the warping and projection to happen at the lowest levels of visual processing in both recognition and generative directions. We thus have eigenfaces (Turk & Pentland, 1991) plus equivalent eigenanalyses for the six other substructures. One can alternatively think of these coefficients as part-specific features of the input.

The middle panel of Figure 1 shows the seven separate sets of coefficients for the particular face, and the three lower panels show how well these coefficients can reconstruct the parts of the face at the different resolutions. The irregular lines show which parts were separately decoded from the coefficients and then pasted together. For the left image, the 4 collections of 20 coefficients representing the individual subparts have been separately decoded and pasted to generate a single image. For the middle image, at an intermediate resolution, the pair of 20 coefficients has been used—one for the two eyes together and one for the nose and mouth together. For the right image, at the lowest resolution, only a single set of coefficients has

been used for all the subparts. The inset images show the principal components and the mean at this lowest resolution. If one looks closely, this reconstruction (depending on only 20 coefficients) is a little worse than that at the high resolution (depending on 80), but the difference is relatively subtle. However, note that the faces are certainly not all the same; for instance, the mean face looks quite different from this particular example.

In total, including all possible resolutions, the complete input associated with each face lives in a $7 \times 20 = 140$ -dimensional space. One way of illustrating the overall task for unsupervised learning is through the covariance matrix of all the faces in this space. That we normalized the dimensionality of each input using PCA implies that there is no off-diagonal structure within the 20×20 blocks along the diagonal of this full covariance matrix, but we can expect substantial structure between the blocks, because of correlations between the features of the subparts (e.g., the eyes are usually similar to each other), and the relationships between the different resolutions. Note that the PCA at the lower resolutions was formally separate from the PCA at the higher resolution, so the coefficients are not trivially related to each other.

The rows of Figure 2 show the top few eigenvectors of the full covariance matrix, ordered by increasing eigenvalue (every fifth one of which is shown on the left of the figure). The eigenvectors can be thought of in 7 sets of 20 columns arising from the image substructures, whence some interrelationships are apparent. For instance, the “rightwards” structure in the eigenvectors arises since PCA was used to generate the fixed-size representations of all seven elements of each complete face description, and the resulting coefficients are also ordered. Also, the forms of the weightings in the components associated with similar parts (e.g., left and right eyes) are somewhat similar, even though separate eigenanalyses were used to generate the 20 coefficients per component.

To capture the common structure among the component coefficients of the faces and their parts, we consider a ϕ -dimensional hidden or latent space ($\phi \leq 140$). That is, we consider the full representation of a face to be a ϕ -dimensional vector \mathbf{h} from which the PCA coefficients \mathbf{x} associated with each of the seven elements can be generated. In terms of relationship 1, this entity should parameterize a map from attentional focus \mathbf{e} to the PCA-reduced, d -dimensional observation \mathbf{x} that could be made at that focus of attention. Following Tenenbaum and Freeman (2000), we do this using a bilinear model, with

$$x_i = \sum_{jk} \mathcal{O}_{ijk} e_j h_k + \eta_i \quad (2.2)$$

where η_i is component-wise independent noise and \mathcal{O} is an observation or imaging tensor that specifies how the latent description \mathbf{h} of the face determines the mapping from attentional focus to observation.

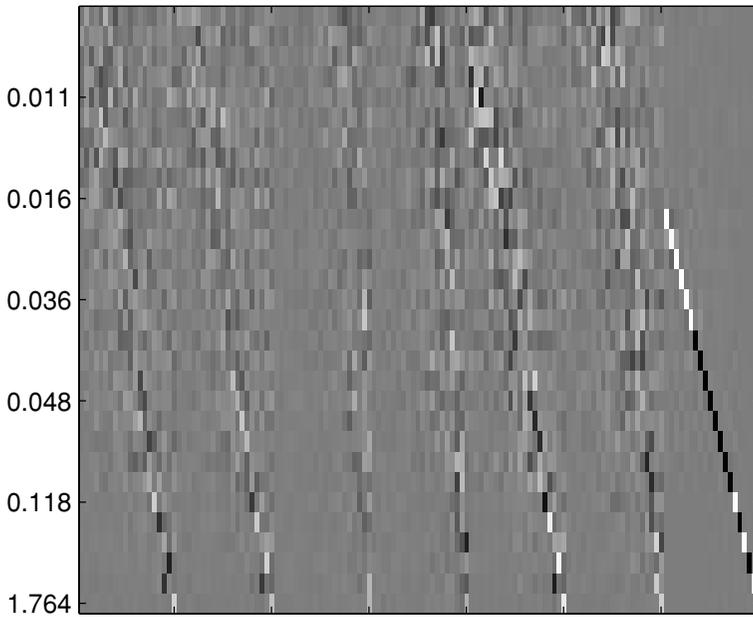


Figure 2: Eigenvectors and eigenvalues of the covariance matrix for the full 140×140 -dimensional representation of the faces. The structure in the 140-dimensional space is evident for the eigenvectors with the largest eigenvalues. The eigenvalues of every fifth eigenvector are shown.

The last required element of the model is to allow for different classes of faces. We do this by considering a mixture model. This adds significant representational power, which is necessary given the highly constrained factor analysis-based representation that we are employing. It can also be seen as an abstraction of the sort of population code representation ubiquitously employed by the cortex. We consider each class (each mixture component) as being a separate (informal) manifold in \mathbf{h} space. We describe the manifold of class c by a ϕ -dimensional mean value \mathbf{v}^c and a $\psi \times \phi$ -dimensional factor loading matrix G^c , where ψ is the true underlying dimension of the manifold. This makes the full latent description of a specific face in this class from the class be

$$\mathbf{h} = \mathbf{g} \cdot G^c + \mathbf{v}^c \quad (2.3)$$

where \mathbf{g} are the (episodic) ψ -dimensional factor values for this specific face and are assumed to have an identity covariance matrix.

Thus, in total, we have the multilinear model

$$x_i = \sum_{jk} \mathcal{O}_{ijk} e_j \left(\sum_l g_l G_{lk}^c + v_k^c \right) + \eta_i. \tag{2.4}$$

We consider the parameters of the imaging model \mathcal{O} to be fixed for all classes of images, since they share a single latent space; the parameters of each class to be G^c and v^c ; and the parameters of a particular face within a class (its unique episodic description) to be the factors \mathbf{g} .

Figure 3 shows the full generative model in the case that there are two classes ($c \in \{1, 2\}$) with a handful of faces assigned (for the moment, arbitrarily) to each. The episodic descriptions of the faces from each class are shown as rectangles containing their ψ -dimensional descriptions (\mathbf{g}). These, via factor loadings (G^c) and together with class-specific mean values v^c (not shown), specify a location in a common ϕ -dimensional space (\mathbf{h}), which acts as the model's hidden representation of the 140-dimensional full representation of the face. The focus of attention (\mathbf{e}) acts in a multilinear manner to select which resolution and which subpart should be imaged. This creates the canonical (in this case, 20-dimensional) representation \mathbf{x} of the part or subpart, which can then be imaged in canonical (warped) coordinates by reversing (as best as possible) the projection from the collection of eigenvectors used to create the reduced input representations.

In the terms of Tenenbaum and Freeman (2000), the imaging process in equation 2.2 is symmetric, with \mathbf{h} and \mathbf{e} being treated equally. Given that there are only a few possible foci of attention, we can also consider an asymmetrical model with $\mathcal{O}_{ik}^e = \sum_j \mathcal{O}_{ijk} e_j$ for the vector \mathbf{e} associated with attentional focus e . In this case, instead of using a ϕ -dimensional mean vector v^c for each class, it is easiest to use a $7d$ -dimensional mean for each class and attentional focus. This has the disadvantage of not capturing the fact that there is coordinated structure in the mean coming from the observation process. However, it has the didactic advantage of making more meaningful the comparisons between different values of the hidden dimension ϕ , uncorrupted by errors in the means. We therefore use this variant in the figures below. Concomitantly, we allow each (of the $e = 1, \dots, 7$) distinct attentional foci to have separate, independent noise terms and so make $\boldsymbol{\eta} \sim N[\mathbf{z}, \mathcal{U}]$, with diagonal covariance matrix

$$\mathcal{U} = \text{diag}(v_1^e \dots v_d^e) \tag{2.5}$$

consisting of the uniquenesses v_i^e (the independent variance terms) for each component i and attentional focus e . In this case, again dropping the class label c for convenience, we can write

$$x_i^e = v_i^e + \sum_k \mathcal{O}_{ik}^e \sum_l g_l G_{lk} + \eta_i^e. \tag{2.6}$$

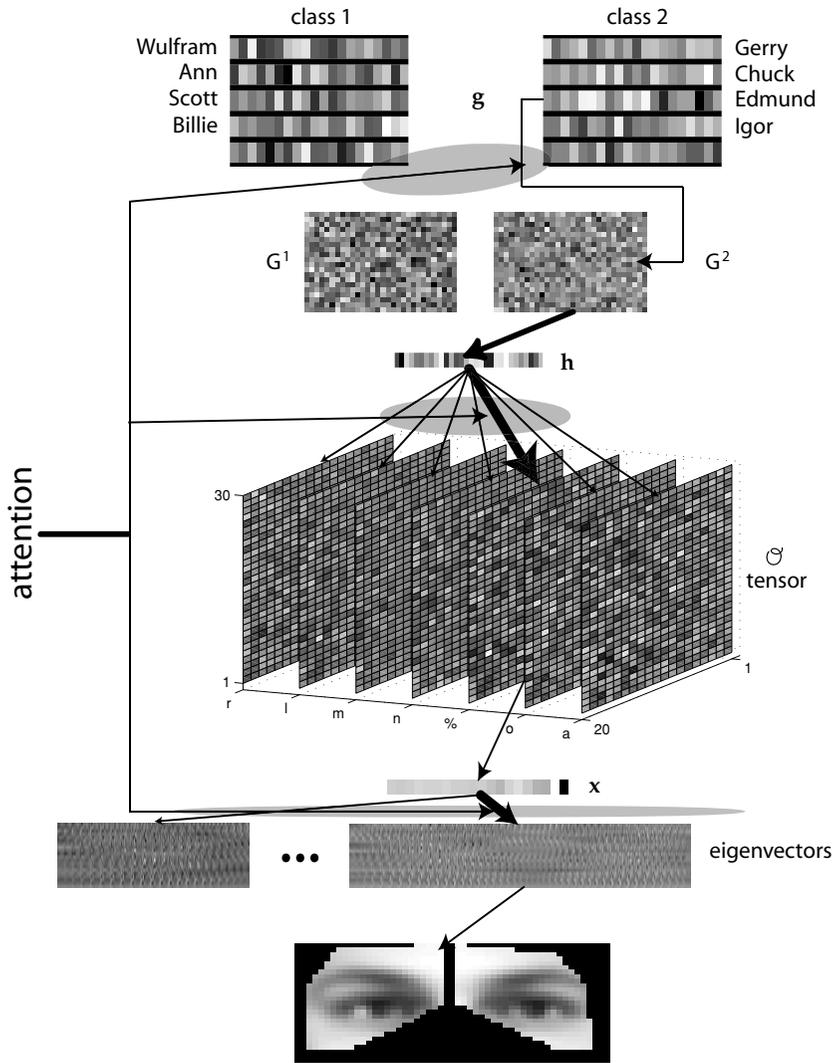


Figure 3: Generative model. Face factors g in one class (2) are mapped through a class-specific factor loading matrix G^2 into a hidden latent representation h and are transformed by the observation tensor \mathcal{O} to give the reduced representation x of one of the subparts, from which the (warped) image can be reconstructed via the principal components. Top-down control (the transparent gray blobs) acts to control the choice of face and the choice of attentional focus, which influences the use of \mathcal{O} and the reconstruction process. The warping is not shown. There is one collection of eigenvectors for each of right eye r , left eye e , nose n , mouth m , both eyes $\%$, mouth and nose o , and the whole face a .

Having specified a rather rich representational structure for the images, we use unsupervised learning to infer the parameters. In section 3, we consider the case that we are ignorant of the true class of each face, turning to the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977). If, however, we do know the classes (in this section, we arbitrarily assign half the faces to one class, the other half to the other), then a maximum likelihood fit of the parameters \mathbf{v}^e , \mathcal{O}^e , G^e , and \mathcal{U} to observed data amounts to a form of weighted least-squares minimization, where the weights arise as part of the full gaussian model. In this section, we consider the related unweighted least-squares problem for which Tenenbaum and Freeman (2000) suggested a solution method involving singular value decomposition in an inner loop. To encompass the next section, in the full weighted problem that has to be solved, we use a conjugate gradient scheme (Carl Rasmussen's *minimize*). As is conventional, we add a baseline to v_i^e to prevent the problem from becoming ill conditioned.

The unweighted case studied by Tenenbaum and Freeman (2000) can be seen as introducing extra parameters \mathbf{g} for each face and then minimizing with respect to \mathbf{v}^e , \mathcal{O}^e , G^e , and \mathbf{g} the mean square error,

$$\left\langle \sum_{ei} \left(v_i^e + \sum_k \mathcal{O}_{ik}^e \sum_l g_l G_{lk} - x_i^e \right)^2 \right\rangle, \tag{2.7}$$

averaged over all the faces in all the classes (which only share \mathcal{O}^e). In this case, we can readily judge the model by considering the reconstructions of the reduced representations x_i ,

$$\hat{v}_i^e + \sum_k \hat{\mathcal{O}}_{ik}^e \sum_l \hat{g}_l \hat{G}_{lk}, \tag{2.8}$$

of the inputs at each attentional focus arising from each face associated with the optimized values.¹

Figure 4 duly shows the result of this optimization in various ways. Figure 4A shows the reconstruction error per pattern as a function of ϕ , the underlying dimension of the hidden space, and ψ , the number of hidden factors. Reconstruction is already quite good for a hidden dimension of around 30 or 40 and around 20 to 30 factors. As might be expected, in the face of multilinearity, it is not generally very useful to trade off ϕ and ψ , the reconstruction is high quality only if both are adequate. This is particularly true for this case of only two classes of face.

Figure 4B shows how the multiple possible observations of a single face are reconstructed as a function of the number of factors ψ , using as a hidden

¹ With the important exception of the prior over the factors, this is very similar to the outcome of a noise clean-up process associated with use of the generative model.

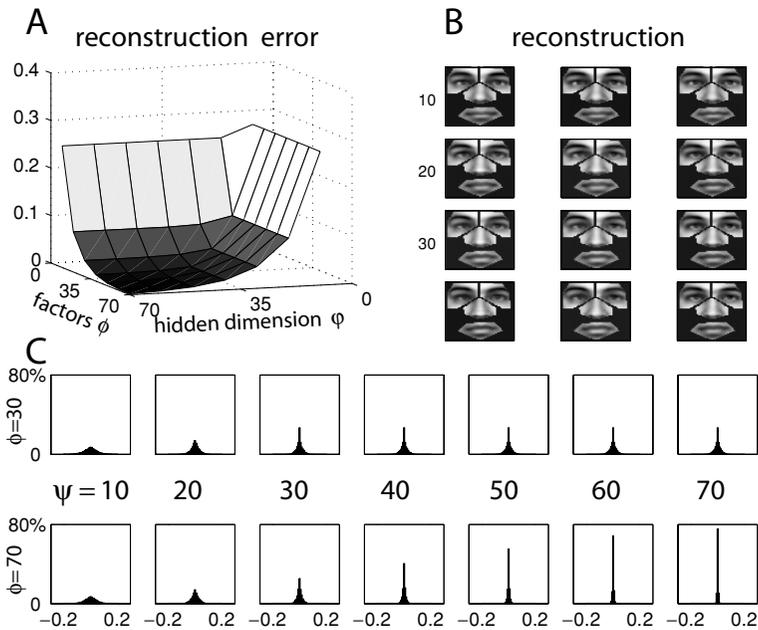


Figure 4: Reconstruction and reconstruction error for the multilinear model. (A) Mean square reconstruction error per full ($7 \times 20 = 140$ -dimensional) representation of the faces from two classes as a function of the dimensionality ϕ of the hidden space and the number ψ of the factors within each space (for comparison, the mean square weight of the representations is 3.7). (B) Reconstruction of a single face pattern (the lowest three images, showing high, medium, and low resolutions as in the bottom row of Figure 1 for $\phi = 30$ dimensions and $\psi = \{10, 20, 30\}$ factors). (C) Histograms of the reconstruction errors for $\phi = 30, 70$ for various numbers of factors.

dimension $\phi = 30$. The model generates the reduced observations $\mathbf{x}^1, \dots, \mathbf{x}^7$, and these have then been mapped into the canonical face coordinates, just as in the bottom row of Figure 1. Again, the differences are rather subtle, and the reconstruction is quite competent even for relatively few factors. This arises because of the redundancy in the full 140-dimensional representation of the faces.

Figure 4C shows the quality of reconstruction in a slightly different manner. Each subplot shows a histogram of the errors in reconstructing all the elements of the \mathbf{x}^i for one whole class of faces. The upper row is for a hidden dimension of $\phi = 30$, the lower for a hidden dimension of $\phi = 70$. Along the rows, the number of factors ψ increases. Again, the high quality of the reconstruction is readily apparent.

A further way to test the model’s ability to capture the structure of the domain is to see how well it can construct one part of a face from other parts. If we know which class the face comes from and the attentional focus of a given sample \mathbf{x} , then we can reconstruct the mean (and variance) of the observations at all the other possible attentional foci. Under gaussian assumptions, the best way to do this is to use the full (in this case, 140×140 -dimensional) covariance matrices shown in Figure 2. Consider the case that we observe a face from its first attentional focus \mathbf{x}^1 . Then write the covariance matrix for the class as

$$\Sigma = \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{1\bar{1}} \\ \hline \Sigma_{11}^T & \Sigma_{1\bar{1}} \end{array} \right),$$

where $\mathbf{1}$ represents all the (in this case, 20) indices associated with the first attentional focus, and $\bar{\mathbf{1}}$ the (120) indices associated with the other attentional foci. In this case, for jointly gaussian $\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^7$, we have the conditional means

$$\mathcal{E}[\mathbf{x}^2, \dots, \mathbf{x}^7 | \mathbf{x}^1] = (\mathbf{x}^1 - \bar{\mathbf{x}}^1) \cdot \Sigma_{11}^{-1} \Sigma_{1\bar{1}} + [\bar{\mathbf{x}}^2, \dots, \bar{\mathbf{x}}^7],$$

where the $\bar{\mathbf{x}}^i$ are the unconditional means of the observations. Remember that the key parts of the reconstructions are therefore the deviations from their means of the reduced representations of the subparts.

Figure 5 shows this for the first class of face. Each small graph shows a histogram of the errors in reconstructing the part shown in the icon in the column from the part shown in the icon in the row. These errors are normalized by the standard deviations of the reconstructed parts so that they are comparable. Various features of these histograms are in accord with obvious intuitions. For instance, given the whole, low-resolution face, the reconstruction of all the other resolutions is good, with the medium resolution easier to reconstruct than the others. The medium-resolution depiction of the combined mouth and nose supports reconstruction of the high-resolution mouth and nose representations much better than it does the high-resolution eye representations, and conversely. The reconstruction of the nose from the eyes or the mouth is superior to the reconstruction of any of the other high-resolution parts.

However, the predictions in Figure 5 are based on the nearly 10^6 components of the class-conditional covariance matrices. We seek reconstruction based on our factor analysis model. Here, given a sample, such as \mathbf{x}^1 , we infer the distribution over the unknown factors \mathbf{g} associated with the whole face, which we can then use to synthesize approximations to $\mathbf{x}^2, \dots, \mathbf{x}^7$. In the next section, we do this (implicitly) using the full factor model; here, we

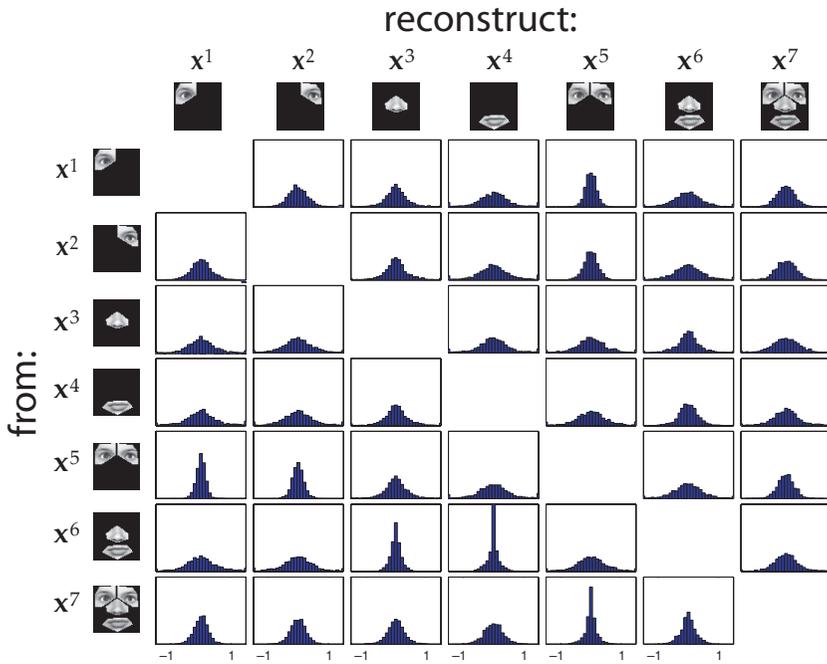


Figure 5: Reconstruction of the reduced representations. The plots show histograms of the errors in reconstruction at high, medium, and low resolutions (columns) based on inputs at each of these resolutions (rows), using the full covariance matrix for the first class of faces. The errors are normalized by the standard deviations of the representations of the reconstructed part to make them directly comparable.

use the solution from the unweighted least-squares problem and therefore an empirical sample factor covariance matrix,

$$\hat{\Gamma}_{ij} = \langle \hat{g}_i \hat{g}_j \rangle, \quad (2.9)$$

averaging over the samples, and uniquenesses,

$$\hat{v}_i^e = \left\langle \left(\hat{v}_i^e + \sum_k \hat{O}_{ik}^e \left(\sum_l \hat{g}_l \hat{G}_{lk} \right) - x_i^e \right)^2 \right\rangle \quad (2.10)$$

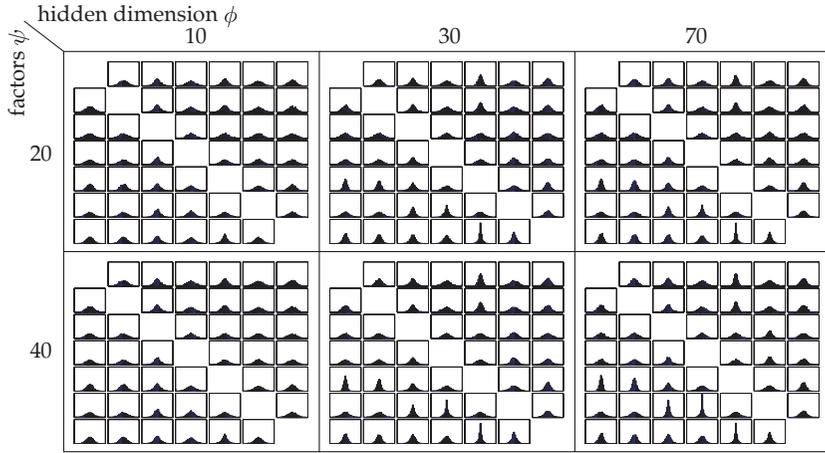


Figure 6: Reconstructions of all the parts from the full trilinear model, using $\psi = 20, 40$ factors (rows) and $\psi = 10, 30, 70$ hidden dimensions (columns). Each plot is exactly as in Figure 5, with the same limits for each graph.

using the best fit \hat{g} to the full x^1, \dots, x^7 . Then if we write the $\psi \times d$ -dimensional matrices,

$$\hat{P}_{li}^e = \sum_k \hat{O}_{ik}^e \hat{G}_{lk},$$

we have, approximately,

$$\mathcal{E}[g_l | \mathbf{x}^1] = (\hat{\Gamma}^{-1} + \hat{P}^1 \cdot [\hat{U}^1]^{-1} \cdot \hat{P}^1)^{-1} \cdot \hat{P}^1 (\mathbf{x}^1 - \hat{\mathbf{v}}^1),$$

where \hat{U}^1 snips out just the uniquenesses associated with the attentional focus that is actually observed. For new attentional focus e , we have

$$\mathcal{E}[x_i^e | \mathbf{x}^1] = \hat{v}_i^e + \sum_k \hat{O}_{ik}^e \left(\sum_l \mathcal{E}[g_l | \mathbf{x}^1] \hat{G}_{lk} \right). \tag{2.11}$$

Figure 6 shows reconstruction (for testing data) according to equation 2.11, using the same format (and the same limits for each individual plot) as Figure 5 and for various values of ϕ and ψ . The first thing to notice is that when there are sufficient factors and dimensions ($\phi = 30; \psi = 20$), reconstruction is nearly indistinguishable from that involving the full covariance matrix (in Figure 5). This is despite the use of many fewer parameters. For

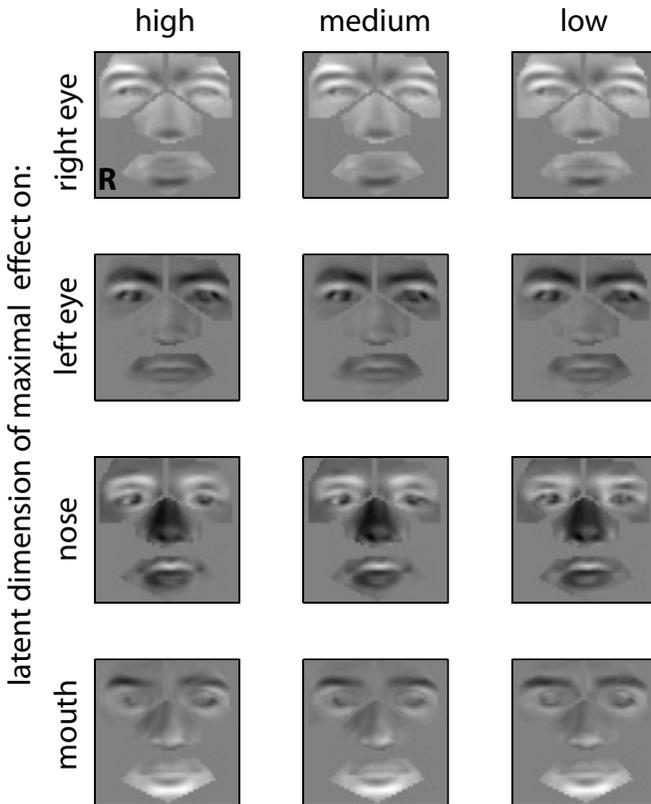


Figure 7: Factor Effects. Each, row shows the effect of a unit change to a single factor on high-, medium-, and low-resolution images, shown in the same format as the bottom row of Figure 1. The four rows are for the four hidden factors that exert the greatest influence over the right and left eyes, nose, and mouth (top to bottom). Influence can be either positive (bright) or negative (dark).

too small a hidden dimension, there is a near-uniform degradation in the quality of all the reconstructions.

As a final view on the multilinear model, we can use the linearity to map the hidden factors back into the image to see the effect of changing one of their values on the whole face. First, we assess which hidden factor exerts the greatest single influence over each of the four high-resolution parts. We then calculate the net effect of changing this factor on all parts of the image and at all resolutions by multiplying together the various observation and factor matrices, and projecting the resulting change back into the full image. Figure 7 shows the result. There is one row each for the factor with the strongest influence over right and left eye, and nose and mouth; each

column shows full images constructed from the subparts at high, medium, and low resolutions, just as in the bottom row of Figure 1.

The most obvious aspect of these plots is that the factors chosen for maximal effect on one subpart do indeed have a greater effect on this subpart than on the others. However, they are much more promiscuous than one might have expected from the reconstruction plots in Figure 6, in which there appeared to be a rather modest effect of one subpart on others. For instance, in these factor effect plots, the changes to the two eyes are almost the same for both factors. There are also more subtle effects—for example, the factor that changes the left eye the most also has a tendency to change the left part of the nose. These plots also show the correct operation of the observation hierarchy in that the changes to parts at the high resolution are replicated at lower resolutions. This was not a forgone conclusion—there are separate parameters for \mathcal{O}^e for different attentional foci e .

3 Clustering

We have so far assumed perfect knowledge of the classes from the outset (using an arbitrary division into two equally sized groups). However, this is clearly unreasonable, and we should also infer the classes from the data themselves. As Tenenbaum and Freeman (2000) noted in their bilinear work, the EM algorithm (Dempster et al., 1977) is ideal for this, provided that we have a fully probabilistic model for each class. In the E step, the posterior probability that each input face comes from each class is assessed. In the M step, these posterior probabilities are used as class-specific importance weightings for each face when both the parameters associated specifically with each class and the common parameters are updated.

In our case, the model of equation 2.4, together with the basic assumptions, amounts to a full generative model (albeit in the eigenfeatures x rather than the pixel input). It is therefore straightforward to perform an exact E step, estimating the posterior responsibilities of the clusters for the data. Here, we consider operating in the regime in which it is possible to study a face from all possible attentional foci in order to calculate the posterior responsibilities. However, the generative model would also make it possible to do incremental learning based on only partial views— x only from a few attentional foci.

Since it is necessary to estimate the uniquenesses, learning is more brittle than for the previous section. We therefore execute only a partial M step, improving the estimates of the parameters of each class given these responsibilities. We also anneal the minimum uniqueness as a way of avoiding premature convergence. Once again, we use Rasmussen's minimize.

Figure 8 shows the result of performing clustering on the entire collection of faces. The faces are relatively homogeneous, and so we do not expect a strong underlying cluster structure. In fact, on synthetic data that actually satisfy the precepts of the full generative model, EM finds the true clusters

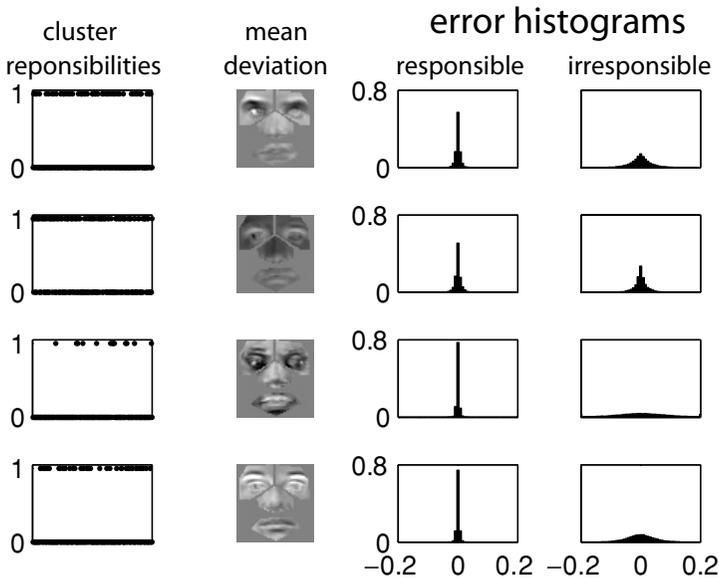


Figure 8: Unsupervised clustering of the face classes. (Column 1) Posterior responsibilities of each of the four clusters for the 190 faces. (Column 2) Deviations of the mean face of each class (those for which the posterior responsibility is greater than 0.8) from the overall mean face. (Column 3) Histograms of the errors in reconstructing the within-class face representations. (Column 4) Histograms of the errors in reconstructing the representations of faces from the other classes. Here $\phi = \psi = 70$. For this figure, the partiality of the M-step involved a fixed number of line searches in minimize for \hat{O} , \hat{G} , and \hat{U} , and a learning rate of 0.01 for changing the prior responsibilities of the clusters. The initial minimal uniqueness was 0.9 and was annealed toward 0.1 at a rate of 0.995 per iteration. We weakened the prior over \mathbf{g} by multiplying the data by a factor of 10. The histograms are directly comparable with those in Figure 4.

(data not shown). The left column shows the posterior responsibilities of each of four classes for all 190 faces. EM does indeed assign faces to each class, though with varying frequencies (27%, 47%, 7%, and 19%, respectively). The second column shows how the mean (warped) face from each class deviates from the overall mean face—some reasonable structure is apparent, such as different overall skin tone. The third and fourth columns show minimal evidence of efficacy of the clustering in that histograms of errors in the reconstructions of all the (reduced representations of the) faces, show that within-class faces (third column) are reconstructed more proficiently than out-of-class faces (fourth column).

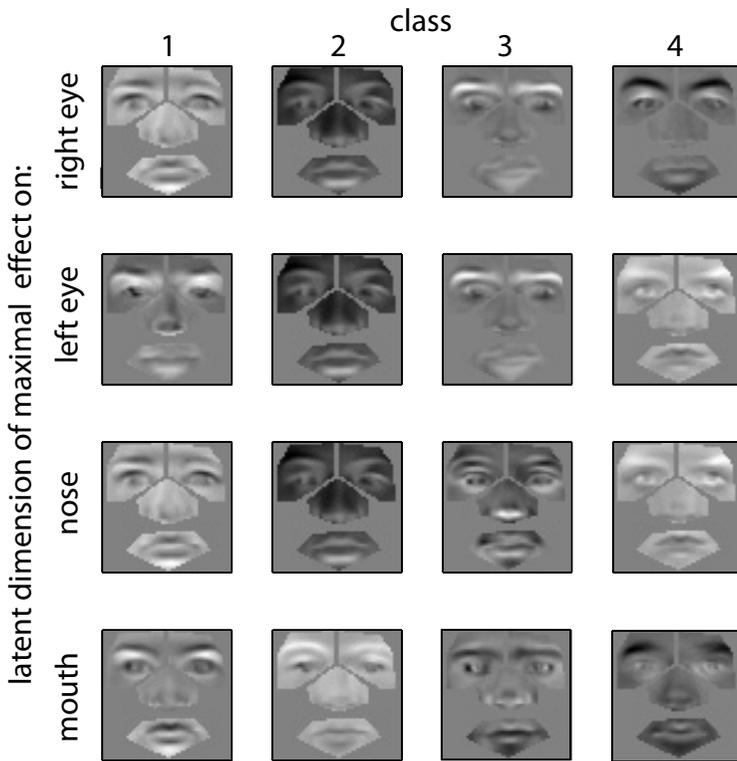


Figure 9: Class-specific factor effects. Each column is associated with one class, each row with the latent factor within the class having the maximal impact on the relevant part. Each figure shows the net effect on high-resolution version of each image of a unit change in the factors. After Figure 7.

Figure 9 shows another view of the differences between the different classes, using the same underlying scheme as in Figure 7. Here, each column is associated with a single class, showing in successive rows the impact on the high-resolution version only (equivalent to just the left-hand column of Figure 7) of the factor with the maximal effect on the right eye, left eye, nose, and mouth. The differences between the different classes are quite marked, despite the fact that all the effects happen through the medium of the same hidden space (\mathbf{h}).

4 Discussion

In this letter, we have considered the representation of hierarchically organized classes of images whose key structuring entity is an explicit variable:

a focus of attention. We showed how to build a factor analysis–based generative model for such classes, and how it can be inferred from data. This was in both a simple case, in which class identity was assumed to be known, and a richer case, involving a mixture-generative model and the EM algorithm, in which unsupervised clustering is also essential. We used the face data from Blanz and Vetter (1999) as our key example. Our prime objective has been to investigate how a single representation can encompass multilevel, statistical, hierarchical structures of different identifiable sorts. Only after we understand this better, perhaps also in richer statistical models than the multilinear gaussian ones here, could the foundation of key cognitively compelling computations over those representations become set. Manipulating more richly structured knowledge is a topic of some current interest in the belief net community (Koller & Pfeffer, 1998; Milch, Marthi, & Russell, 2004); we have considered it in more connectionist terms.

Our work has a rather diverse range of links. First, it took its structuring of the problem of hierarchical representation from Riesenhuber and Dayan (1996). That article set out to put into context the neurophysiological results of Connor et al. (1996), who tested Olshausen et al.'s (1993) shifter model of attention. Connor et al. found that a major effect of specifying the focus of (visual) attention was not to translate and scale the mapping from lower to higher visual areas, as expected from the shifter model, but rather to scale multiplicatively the activity of at least one population of V4 neurons. Riesenhuber and Dayan (1996) and Salinas and Abbott (1997) treated this scaling as part of a basis function representation involving the simultaneous coding of the focus of attention and image-object features. As has been well explored (Poggio, 1990), particularly through the medium of models of parietal cortex (Pouget & Sejnowski, 1997; Deneve & Pouget, 2003), such basis function mappings allow a simple instantiation of complex functions of all the variables represented. One can see the multilinear form in equation 2.6 in basis function terms, involving an interaction between a representation $\sum_l g_l G_{lk}$ of the image contents and the effect \mathcal{O}_{ik}^e of the attentional focus e . More general basis functions could be used to allow nonlinear models of the classes themselves and of the effects of the focus of attention.

Amit and Mascaro's (2003) shifter-like model is also related. That model has an attractively sophisticated shifting process that integrates bottom-up and top-down information; it would be interesting to employ within it the sort of hierarchical representations of the top-down information that has been our focus. We have relied on shifting to achieve the sort of preprocessing normalization leading to our observations, x .

A second key antecedent is the work of Tenenbaum and Freeman (2000) on multilinear generative models for the mostly unsupervised separation of different factors ("content," for us the nature of the face, and "style," the attentional focus) that interact to determine inputs. They articulated a general framework to study the sort of multiplicative interactions that we have employed, related them to a range of existing ideas about bilinearity

in perceptual domains (Koenderink and van Doorn, 1997), and showed a most interesting application to typefaces. The particular method that Tenenbaum and Freeman (2000) used to fit their generative model (avoiding one gradient step through the use of singular value decomposition) could be adapted to our case, but an iterative method seems to be required for trilinear and higher-order models in any event. We used a gradient-based minimizer to solve the whole problem.

Vasilescu and Terzopoulos (2002, 2003), building partly on the work of De Lathauwer (1997) and Kolda (2001) (and based on ideas dating back at least to Tucker, 1966), used a tensor extension to singular valued decomposition (SVD) to find what can be seen as a joint coordinate scheme for structured collections of images. Take the case of faces. Their method starts from a data tensor, with the different dimensions of structural variation of the images (such as viewpoints, lighting conditions, and identities) kept as separate dimensions. Just as SVD on a normal two-dimensional matrix finds left and right coordinate systems for the two spaces acted on by the matrix, together with singular values that link them, joint SVD on the tensor finds coordinate systems for each dimension together with what is called a core tensor that links them collectively. Each coordinate system parameterizes its dimension of variability. In terms of this scheme, our method is rather like using the focus of attention as the independent dimension and marginalizing over identity (which can then be separated through the medium of the mixture model). In these terms, we might expect the observation tensor \mathcal{O} to have a formal relationship with the SVD coordinates associated with the focus of attention. Given this, extensions of the tensor decomposition idea, such as to independent components analysis (De Lathauwer & Vandewalle, 2004; Vasilescu & Terzopoulos, 2005), could be most useful directions for our work on representation.

A third link is to tensor product-based representations of structured knowledge (Smolensky, 1990; Plate, 1995, 2003; Gayler, 1998). This strand of work has placed most of its efforts into the problem of the representation of arbitrary episodic structured facts, with representational elements newly minted for each new case. The same is true for methods that are further from tensor product notions, such as Rachkovskij and Kussul's (2001) context-dependent thinning method for binary distributed representations or Kanerva's (1996) binary spatter codes. By contrast, we have focused on the semantic structure underlying domains such as faces. However, the basic linear operations inherent in the multilinear models (such as equation 2.4) are indeed just tensor products of various sorts.

Perhaps an even closer link is to recursive autoassociative memories (RAAMs; Pollack, 1990), their ancestor in Hinton's notion of reduced descriptions (Hinton, 1990) and their relational descendants (Sperduti, 1994), since at their heart they are autoencoders, which are best understood as forms of statistical generative model. However, again, RAAMs are normally considered in episodic rather than semantic terms, so the influence

exerted by the overall statistical structure of a domain can be hard to discern. Paccanaro and Hinton's (2001) linear relational embedding (LRE) can be seen as an intermediate case. LRE, which significantly generalizes and formalizes Hinton's (1989) famous family trees network, does learn aspects of the semantics of a domain, considering the overall structure of the group of related facts about a number of different episodic examples.

A final important link arises through ideas in computational vision for the representation of structured objects such as faces. There is a huge wealth of techniques based on generative models of various sorts, from the sort of image-based methods favored by Edelman (1999) through a variety of approaches that decompose objects into parts and learn something about the relative positions and form of these parts. For some methods, the parts are sometimes intended to capture something about the true structure of the object (Fischler & Elschlager, 1973; Grenander, 1976–1981; Mjolsness, 1990; Revow, Williams, & Hinton, 1996). For other methods, the parts are features more like dense subcomponents, or local patches of the images, or local wavelet coefficients (Burl et al., 1995, 1998; Schiele & Crowley, 1996, 2000; Fei-Fei et al., 2003; Liebe & Schiele, 2003, 2004; Schneiderman & Kanade, 2004; Amit & Trouvé, 2005). Most, though not all, methods incorporate explicit knowledge about the geometrical relationships among the parts and have it play a key role in the recognition processes of detection and classification. Our method is best seen as image based, and although it has implicit information about these relationships in its ability to generate representations of parts from wholes (one of the main intents in Riesenhuber & Dayan, 1996), we have not considered such sophisticated recognition issues.

The most important future direction for this work is in the direction of knowledge structures that are more general than images. Take stories as an extreme, but seductively motivating, example, to which, for instance, Dolan (1989) took the idea of tensor product representations. However, as with other tensor product notions, this was formulated before the widespread formulation of the sort of sophisticated statistical unsupervised learning model that Tenenbaum and Freeman (2000) promulgated. Stories of a given class (just like faces of a given class) share a semantic structure that defines constraints among the actors and actions in the story (just like the eyes in a face). A most critical difference is that although there are intuitive notions of scale (perhaps summarization scale) and substructure in stories, there is no obvious equivalent of what we have called the attentional focus \mathbf{e} , as a way of defining observations at different scales or resolutions. One possible generalization of the key mapping definition (see equation 2.1) is:

$$\begin{aligned} \text{story : question} &\Rightarrow \text{answer} \\ \mathcal{I} : \quad \quad \quad \mathbf{e} &\Rightarrow \mathbf{x} \end{aligned} \tag{4.1}$$

with *question* and *answer* being coded in the same latent space. In a linear version, this would imply a generalization of equation 2.4, with

$$x_i = \sum_{jk} \mathcal{O}_{ijk} \left(\sum_m q_m \mathcal{Q}_{mj} \right) \left(\sum_l g_l \mathbf{G}_{lk}^c \right) + \eta_i \quad (4.2)$$

(ignoring the means), where \mathbf{q} are the hidden factors underlying the *question*, \mathbf{x} is a representation of the *answer*, and \mathcal{O} maps together the question and story representations. Altogether $\mathbf{q} \cdot \mathcal{Q}$ acts as the equivalent of the attentional focus \mathbf{e} . In this case, the *answer* should perhaps live in the same representational space as the *question*, that is, be itself captured through factors \mathcal{Q} , although this poses a rather more challenging unsupervised learning problem.

Of course, the restriction to a purely multilinear generative model is rather severe for learning. It will be important to consider nonlinear generalizations of this in which the eye position and the latent space (or the *question* and the *story*) interact in richer manner. Some of the recent structured image representations mentioned above may provide some pointers. The first of a very large number of steps might be to take advantage of the much greater flexibility of a mixture model.

Acknowledgments

I am very grateful to Max Riesenhuber for most helpful discussion and comments and particularly for encouraging and helping me to use the faces from Blanz and Vetter (1999). Two anonymous reviewers and Yali Amit and Chris Williams made very helpful suggestions. I also thank Volker Blanz and Thomas Vetter for making the faces available, and particularly to the latter for allowing me to use the fiducial markers. Support was from the Gatsby Charitable Foundation.

References

- Amit, Y., & Mascaro, M. (2003). An integrated network for invariant visual detection and recognition. *Vision Research*, *43*, 2073–2088.
- Amit, Y., & Trouvé, A. (2005). *POP: Patchwork of parts models for object recognition*. Unpublished technical report. Available online from <http://galton.uchicago.edu/amit/Papers/pop.pdf>.
- Beymer, D., & Poggio, T. (1996). Image representations for visual learning. *Science*, *272*, 1905–1909.
- Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. In *SIGGRAPH'99* (pp. 187–194). New York: ACM Computer Society Press.
- Bridgeman, B., van der Heijden, A. H. C., & Velichkovsky, B. (1994). Visual stability and saccadic eye movements. *Behavioral and Brain Sciences*, *17*, 247–258.

- Burl, M. C., Leung, T. K., & Perona, P. (1995). Face localization via shape statistics. In M. Bichsel (Ed.), *Proceedings of the International Workshop on Automatic Face and Gesture Recognition* (pp. 154–159). Piscataway, NJ: IEEE.
- Burl, M., Weber, M., & Perona, P. (1998). *A probabilistic approach to object recognition using local photometry and global geometry*. Berlin: Springer.
- Connor, C. E., Gallant, J. L., Preddie, D. C., & Van Essen, D. C. (1996). Responses in area V4 depend on the spatial relationship between stimulus and attention. *Journal of Neurophysiology*, *75*, 1306–1308.
- Crandall, D., Felzenszwalb, P., & Huttenlocher, D. (2005). Spatial priors for part-based recognition using statistical models. *Proceedings of the International Conference of Computer Vision and Pattern*. Piscataway, NJ: IEEE.
- Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The Helmholtz machine. *Neural Computation*, *7*, 889–904.
- De Lathauwer, L. (1997). *Signal processing based on multilinear algebra*. Unpublished doctoral dissertation, Katholieke Universiteit Leuven, Belgium.
- De Lathauwer, L., & Vandewalle, J. (2004). Dimensionality reduction in higher-order signal processing and rank- (R_1, R_2, \dots, R_N) reduction in multilinear algebra. *Linear Algebra Applications*, *391*, 31–55.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B39*, 1–38.
- Deneve, S., & Pouget, A. (2003). Basis functions for object-centered representations. *Neuron*, *37*, 347–359.
- Dolan, C. P. (1989). *Tensor manipulation networks: Connectionist and symbolic approaches to comprehension, learning and planning* (Tech. Rep. UCLA-AI-89-06). Los Angeles: Computer Science Department, AI Lab, UCLA.
- Edelman, S. (1999). *Representation and recognition in vision*. Cambridge, MA: MIT Press.
- Fei-Fei, L., Fergus, R., & Perona, P. (2003). A Bayesian approach to unsupervised oneshot learning of object categories. In *Proceedings of the International Conference on Computer Vision ICCV* (Vol. 1). Piscataway, NJ: IEEE.
- Fergus, R., Perona, P., & Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. In *Proceedings of the International Conference of Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE.
- Fischler, M. A., & Elschlager, R. A. (1973). The representation and matching of pictorial structures. *IEEE Transactions on Computers*, *22*, 67–92.
- Gayler, R. W. (1998). Multiplicative binding, representation operators and analogy. In K. Holyoak & D. Gentner (Eds.), *Advances in analogy research: Integration of theory and data from the cognitive, computational, and neural sciences*. Sofia, Bulgaria: New Bulgarian University.
- Grenander, U. (1976–1981). *Lectures in pattern theory I, II and III: Pattern analysis, pattern synthesis and regular structures*. Berlin: Springer-Verlag.
- Grimes, D. B., & Rao, R. P. N. (2005). Bilinear sparse coding for invariant vision. *Neural Computation*, *17*, 47–73.
- Hinton, G. E. (1989). Learning distributed representations of concepts. In R. G. M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology* (pp. 46–61). New York: Oxford University Press.

- Hinton, G. E. (1990). Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence*, 46, 47–76.
- Hinton, G. E. (Ed.). (1991). *Connectionist symbol processing*. Cambridge, MA: MIT Press.
- Hinton, G. E., Dayan, P., & Revow, M. (1997). Modeling the manifolds of images of handwritten digits. *IEEE Transactions on Neural Networks*, 8, 65–74.
- Hinton, G. E., & Ghahramani, Z. (1997). Generative models for discovering sparse distributed representations. *Philosophical Transactions of the Royal Society of London*, B, 352, 1177–1190.
- Hinton, G. E., & Zemel, R. S. (1994). Autoencoders, minimum description length and Helmholtz free energy. In J. D. Cowan, G. Tesauro, & J. Altspector (Eds.), *Advances in neural information processing systems*, 6 (pp. 3–10). San Mateo, CA: Morgan Kaufmann.
- Kanerva, P. (1996). Binary spatter-coding of ordered K-tuples. In C. von der Malsburg, W. von Seelen, J. C. Vorbrüggen, & B. Sendhoff (Eds.), *Proceedings of ICANN 1996* (pp. 869–873). Berlin: Springer-Verlag.
- Koenderink, J. J., & van Doorn, A. J. (1997). The generic bilinear calibration-estimation problem. *International Journal of Computer Vision*, 23, 217–234.
- Kolda, T. G. (2001). Orthogonal tensor decompositions. *SIAM Journal on Matrix Analysis and Applications*, 23, 243–255.
- Koller, D., & Pfeffer, A. (1998). Probabilistic frame-based systems. In *Proceedings of the 15th National Conference on Artificial Intelligence* (pp. 580–587). Madison, WI: AAAI Press.
- Liebe, B., & Schiele, B. (2003). Interleaved object categorization and segmentation. In *British Machine Vision Conference (BMVC'03)*. Norwich, UK.
- Liebe, B., & Schiele, B. (2004). Scale invariant object categorization using a scale-adaptive mean-shift search. In *DAGM'04 Annual Pattern Recognition Symposium* (pp. 145–153). Berlin: Springer-Verlag.
- Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, 21, 105–117.
- MacKay, D. M. (1956). The epistemological problem for automata. In C. E. Shannon & J. McCarthy (Eds.), *Automata studies* (pp. 235–251). Princeton, NJ: Princeton University Press.
- Milch, B., Marthi, B., & Russell, S. (2004). BLOG: Relational modeling with unknown objects. In *Proceedings of the ICML-04 Workshop on Statistical Relational Learning*. Banff, Canada: International Machine Learning Society.
- Mjolsness, E. (1990). *Bayesian inference on visual grammars by neural nets that optimize* (Tech. Rep. YALEU/DCS/TR-854). New Haven, CT: Computer Science Department, Yale University.
- Mumford, D. (1994). Neuronal architectures for pattern-theoretic problems. In C. Koch & J. Davis (Eds.), *Large-scale theories of the cortex* (pp. 125–152). Cambridge, MA: MIT Press.
- Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.
- Olshausen, B. A., Anderson, C. H., & Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, 13, 4700–4719.
- Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381, 607–609.

- Paccanaro, A., & Hinton, G. E. (2001). Learning distributed representation of concepts using linear relational embedding. *IEEE Transactions on Knowledge and Data Engineering*, *13*, 232–245.
- Plate, T. A. (1995). Holographic reduced representations. *IEEE Transactions on Neural Networks*, *6*, 623–641.
- Plate, T. A. (2003). *Holographic reduced representations*. Stanford, CA: CSLI Publications.
- Poggio, T. (1990). A theory of how the brain might work. *Cold Spring Harbor Symposium on Quantitative Biology*, *55*, 899–910.
- Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence*, *46*, 77–105.
- Pouget, A., Dayan, P., & Zemel, R. S. (2000). Computation with population codes. *Nature Reviews Neuroscience*, *1*, 125–132.
- Pouget, A., & Sejnowski, T. J. (1997). Spatial transformations in the parietal cortex using basis functions. *Journal of Cognitive Neuroscience*, *9*, 222–237.
- Rachkovskij, D. A., & Kussul, E. M. (2001). Binding and normalization of binary sparse distributed representations by context-dependent thinning. *Neural Computation*, *13*, 411–452.
- Rao, R. P. N., Olshausen, B. A., & Lewicki, M. S. (2002). *Probabilistic models of the brain*. Cambridge, MA: MIT Press.
- Revow, M., Williams, C. K. I., & Hinton, G. E. (1996). Using generative models for handwritten digit recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *18*, 592–606.
- Riesenhuber, M., & Dayan, P. (1996). Neural models for part-whole hierarchies. In M. C. Mozer, M. I. Jordan, & T. Petsche (Eds.), *Advances in neural information processing systems*, *9* (pp. 17–23). Cambridge, MA: MIT Press.
- Riesenhuber, M., Jarudi I., Gilad, S., & Sinha, P. (2004). Face processing in humans is compatible with a simple shape-based model of vision. *Proceedings of the Royal Society of London, B (Suppl.)*, *271*, S448–S450.
- Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *11*, 1019–1025.
- Salinas, E., & Abbott, L. F. (1997). Invariant visual responses from attentional gain fields. *Journal of Neurophysiology*, *77*, 3267–3272.
- Schiele, B., & Crowley, J. L. (1996). Probabilistic object recognition using multidimensional receptive field histograms. In *International Conference on Pattern Recognition*. Piscataway, NJ: IEEE.
- Schiele, B., & Crowley, J. L. (2000). Recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision*, *36*, 31–50.
- Schneiderman, H., & Kanade, T. (2004). Object detection using the statistics of parts. *International Journal of Computer Vision*, *56*, 151–177.
- Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*, 159–216.
- Sperduti, A. (1994). Labeling RAAM Connection Science, *6*, 429–459.
- Sudderth, E. B., Torralba, A., Freeman, W. T., & Willsky, A. S. (2005). Learning hierarchical models of scenes, objects and parts. In *Proceedings of the International Conference on Computer Vision*. Piscataway, NJ: IEEE.

- Tenenbaum, J. B., & Freeman, W. T. (2000). Separating style and content with bilinear models. *Neural Computation*, *12*, 1247–1283.
- Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, *31*, 279–311.
- Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, *3*, 71–86.
- Vasilescu, M. A. O., & Terzopoulos, D. (2002). Multilinear analysis of image ensembles: TensorFaces. In A. Heyden, G. Sparr, M. Nielsen, & P. Johansen (Eds.), *Computer vision* (pp. 447–460). Berlin: Springer-Verlag.
- Vasilescu, M. A. O., & Terzopoulos, D. (2003). Multilinear subspace analysis for image ensembles. In *Proceedings of Computer Vision and Pattern Recognition Conference, CVPR '03* (Vol. 2, pp. 93–99). Piscataway, NJ: IEEE.
- Vasilescu, M. A. O., & Terzopoulos, D. (2005). Multilinear independent components analysis. In *Proceedings of Computer Vision and Pattern Recognition Conference, CVPR '05* (Vol. 2, pp. 547–553). Piscataway, NJ: IEEE.
- Vetter, T., & Poggio, T. (1997). Linear object classes and image synthesis from a single example image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*, 733–742.
- von der Malsburg, C. (1988). Pattern recognition by labelled graph matching. *Neural Networks*, *1*, 141–148.

Received August 1, 2005; accepted April 4, 2006.