1997

A comparison of local versus global image decompositions for visual speechreading

Michael S. Gray[1,3], Javier R. Movellan[1], Terrence J. Sejnowski[2,3]
Departments of Cognitive Science[1] and Biology[2]
University of California, San Diego
La Jolla, CA 92093
and
Howard Hughes Medical Institute[3]
Computational Neurobiology Lab
The Salk Institute, P. O. Box 85800
San Diego, CA 92186-5800
Email: mgray, jmovellan, tsejnowski@ucsd.edu

Abstract

What is the appropriate spatial scale for image representation? In the primate visual system, receptive fields are small at early stages of processing (area V1), and larger at late stages of processing (areas MT, IT). In the current work, we explore the efficiency of local and global image representations on an automatic visual speech recognition task using an HMM as the recognition system. We compare local and global principal component and independent component image representations for the task. Local representations consistently and significantly outperformed global representations in terms of generalization to new speakers.

1 INTRODUCTION

Global principal component representations have been widely explored for image processing tasks (Turk & Pentland [9]). In this representation, one typically obtains image bases which are non-local (i.e., each image in the basis set has non-zero energy distributed about the whole image). Recently it has been proposed that decompositions that use a local image basis are preferable (Olshausen & Field [6]; Penev & Atick [8]). Indeed the receptive fields of neurons in early stages of the primate visual system are local, suggesting that the visual input is decomposed using sets of shifted local basis images. However, the advantages and disadvantages of local versus global decompositions have seldom been studied in machine vision tasks. The main goal of this paper is to compare local versus global image decomposition techniques on an automatic visual speechreading task. In addition we compare decompositions obtained with principal component analysis versus those obtained with independent component analysis (Bell & Sejnowski [2, 3]). The different techniques are analyzed both qualitatively for spectral characteristics, and quantitatively for their performance on a visual speechreading task.

2 METHODS AND MODELS

The task consisted of automatic recognition of the first four digits in English using the Tulips1 database (Movellan [5]): 96 digitized movies of 12 undergraduate students (9 males, 3 females) from the Cognitive Science Department at UC-San Diego. The images were normalized so that variations due to translation, scale, and planar rotation were eliminated. This normalization was done using parameters obtained from active shape contour modeling by Luettin, Thacker, and Beet [4] on the same database. The dataset consisted of 934 images, each one 65 pixels × 87 pixels (5655 total) in size. Sample images are shown in the left panels of figure 1.

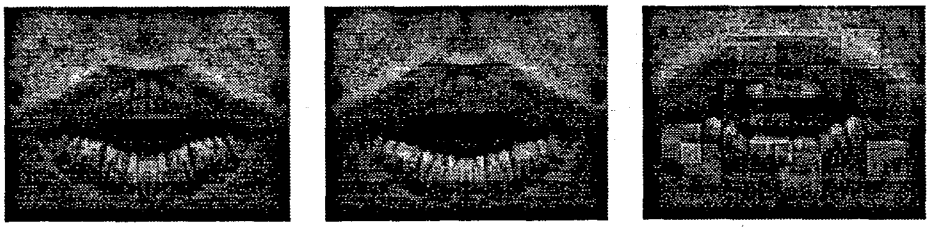

Figure 1: Left two panels: Successive video frames (frames 1 and 2) from a subject saying the digit "one". These images have been made symmetric by averaging left and right pixels relative to the vertical midline. Right panel: Random lip patches (12 pixels x 12 pixels) used to develop local PCA and local ICA filters.

2.1 GLOBAL DECOMPOSITIONS

GLOBAL PCA Principal component analysis (PCA) is an unsupervised procedure that provides an orthonormal basis for the images in the database. The basis images are the eigenvectors of the covariance matrix whose (i, j) entry is the empirical covariance between the i^{th} and j^{th} pixels across all the images in the database. The obtained image bases are ordered according to the eigenvalues of the covariance matrix. It can be shown that the coordinates of the images in the new basis are uncorrelated. Moreover, the variance of each coordinate equals the eigenvalues of the corresponding eigenimages (eigenvectors). For natural image datasets, the distribution of eigenvalues is typically exponential. Moreover, the image bases obtained using PCA are typically non-local in the space domain (i.e., have non-zero energy distributed over the whole image).

The top row of figure 2 shows the first 5 eigenimages in our database. Their magnitude spectrum is shown in the second row. These first 5 eigenimages have most of the energy localized in low spatial frequencies. This localization in the frequency domain reflects the lack of localization in the space domain. We investigated the performance of image representations that used the coordinates of each image in the database with respect to the first 50 eigenimages. These coordinates accounted for 94.5% of the variance in the images. In addition, the delta coordinates (the difference in PCA projection from one time step to the next) were used in the representation at each time step.

GLOBAL ICA The goal of independent component analysis (ICA) is to perform a nonlinear monotonic transformation of N random variables such that the transformed variables are statis-

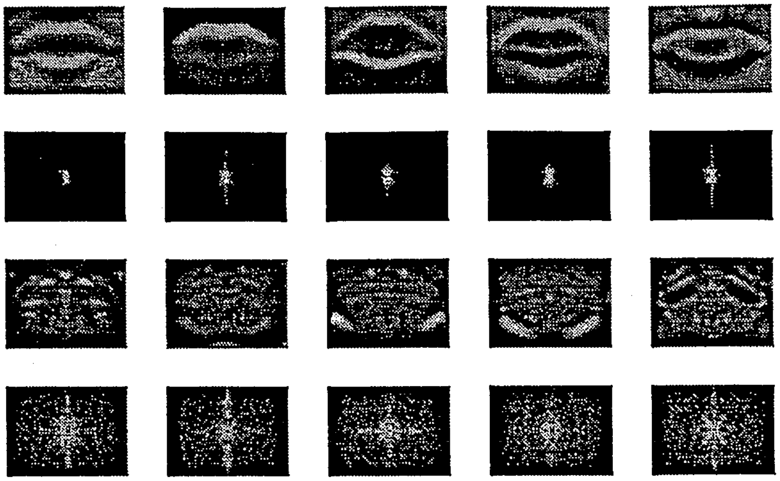

Figure 2: Row 1: Global principal component filters ordered with first eigenimage on left. Row 2: Magnitude spectrum of global principal components. Row 3: Global independent component filters ordered according to magnitude (shortest filter on left). Row 4: Magnitude spectrum of global independent components.

tically independent, not just uncorrelated. The approach proposed in (Bell & Sejnowski [3]) is to maximize the joint entropy of output random variables, where each output random variable is a linear combination of the input variables followed by a logistic nonlinearity. If the outputs are truly independent, then the mutual information between each pair of outputs at the end of training is zero: $I(U_i, U_j) = 0, \forall i \neq j$. Note that mutual information depends on statistics of all orders. PCA, on the other hand, only considers second-order statistics (decorrelation). Moreover, PCA provides an orthonormal basis, whereas the ICA basis images are not necessarily orthogonal.

In PCA, the eigenvalues indicate the proportion of variance accounted for by projections onto each eigenimage. For ICA, there is no obvious analogue to this measure. Bell (1997, personal communication) has suggested that the magnitude (L_2 norm) of the filter is inversely proportional to the variance of the original source. In other words, the filter with the lowest magnitude corresponds to the independent source that is most strongly present in the original data.

The goal of ICA learning is to find the unmixing matrix $\mathbf{W}$ that recovers the independent sources $\mathbf{U}$ from the image dataset $\mathbf{X}$: $\mathbf{U} = \mathbf{WX}$, where $\mathbf{X}$ is 934 images × 5655 pixels (65 rows × 87 columns), $\mathbf{W}$ is 934 × 934, and $\mathbf{U}$ is the same size as $\mathbf{X}$. The weight changes in $\mathbf{W}$ are proportional to the gradient of the entropy of the outputs $\mathbf{O} = g(\mathbf{U})$, where $g(\cdot)$ is an invertible bounded nonlinear transfer function (e.g., sigmoid). However, the large number of images in our database makes this direct application of ICA infeasible because of computation time. We used the following approach, inspired by work of Bartlett and Sejnowski [1]. Instead of the original 934 × 5655 data matrix, the first 50 eigenimages were used as input — one eigenimage per row, for a matrix $\mathbf{E}$ of size 50 × 5655. The independent components of this eigenimage representation $\mathbf{E}$ were then found by learning the unmixing matrix $\mathbf{W}$: $\mathbf{U} = \mathbf{WE}$. The first 5 independent components (smallest magnitude) are shown in the third row of figure 2. The fourth row shows their magnitude spectrum. The independent components have a broader frequency spectrum than the principal components, and thus they are also more local in the space domain. To construct the ICA representation for the *original* lip images $\mathbf{X}$ (not the eigenimages $\mathbf{E}$ on which ICA was trained) we solve for $\mathbf{Y}$, a linear combination of the independent sources $\mathbf{U}$ which reconstructs the original dataset, $\mathbf{X}$:

$$\mathbf{Y}\,\mathbf{U} = \mathbf{X} \tag{1}$$

$$\mathbf{Y} = \mathbf{X}\,\mathbf{U}^{\ddagger} \tag{2}$$

where $\mathbf{U}^{\ddagger}$ is the pseudo-inverse of $\mathbf{U}$. The matrix $\mathbf{Y}$, combined with the delta coordinates at each time step, served as the global ICA representation of the lip images.

2.2 LOCAL DECOMPOSITIONS

The basis images obtained with PCA and ICA were global for the most part. To test the idea that a local basis is better, we computed PCA and ICA filters while enforcing a local constraint. Small image patches (12 pixel × 12 pixel) were chosen from random locations in the lip images. Twenty patches were collected from each of the 934 images in the dataset for a total of 18680 patches. A sample of these random patches (superimposed on a lip image) is shown in the right panel of figure 1. The obtained matrix of local lip patches (144 pixels × 18680 patches) formed the dataset for the local PCA and ICA analysis. Hereafter we refer to the 12 pixel × 12 pixel images obtained via PCA or ICA as "local kernels". The basis images were obtained by centering a local kernel obtained with PCA or ICA onto a particular location of a 65 × 87 matrix and padding the rest of the matrix with zeros. The center of the local kernel was shifted into 300 different locations uniformly distributed across the 65 × 87 possible locations, for a total of 300 basis images. Hereafter we refer to the 300 shifted images as the basis induced by a local PCA or ICA filter. Note that this procedure forces the basis images to be local in the space domain. A sample basis image is shown in the left panel of figure 3.

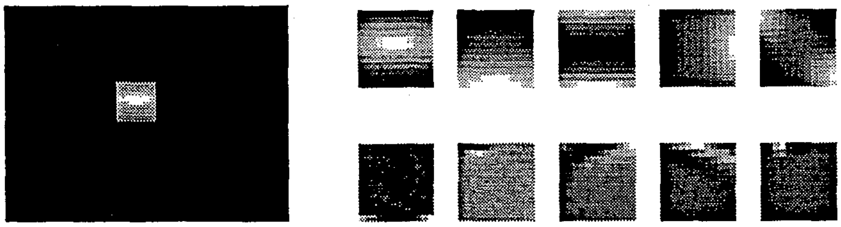

Figure 3: Left: Sample local image basis. Upper right: Local principal component filters ordered with first eigenpatch on left. Lower right: Local independent component filters ordered according to magnitude (shortest filter on left).

LOCAL PCA The upper right panels in Figure 3 show the first 5 local eigenimages (eigenpatches). In practice, to obtain the projection of a lip image onto each of the 300 image bases induced by an eigenpatch, we filtered the lip image using the eigenpatch as the point spread function and subsampled the filtered image at 300 uniformly distributed locations (15 locations vertically by 20 locations horizontally). The delta representation was formed from the difference in the filter responses at the same location between subsequent time frames. Because the lip images were horizontally symmetric, only half of the filtered and delta images were used, resulting in a 300-dimensional input vector.

LOCAL ICA Each image patch $\mathbf{x}$ is assumed to have been formed through the weighted linear combination of a set of basis functions (Olshausen & Field [6]; Bell & Sejnowski [3]). The task of ICA is to find the weight matrix $\mathbf{W}$ (144 × 144) that recovers the basis functions $\mathbf{U}$ (144 × 18680)

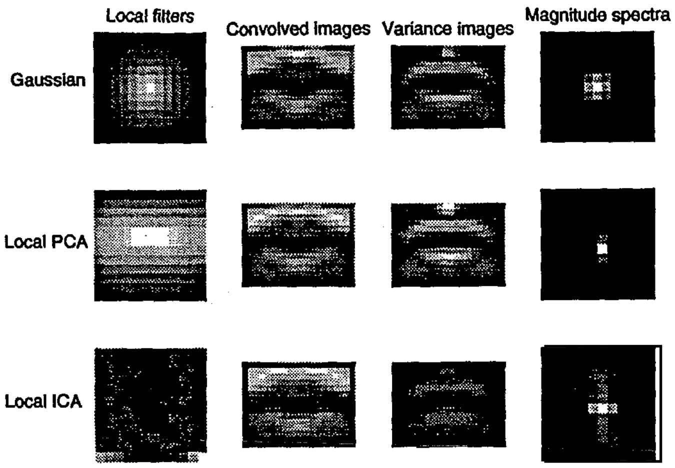

Figure 4: Column 1: Local filters in pixel space. Column 2: Sample lip image convolved with local filters and downsampled. Column 3: Variance of local filter output as a function of pixel location. Column 4: Magnitude spectrum of local filters.

from the data **X** (144 × 18680): **U** = **WX**. The rows of **W** are the local kernels which recover these basis functions.

In a manner identical to PCA, the five lowest magnitude ICA filters (Figure 3, lower right panels) were each used to process the lip images. The images were filtered, then downsampled to 15 × 20 pixels. Delta representations were formed from differences in filter outputs between subsequent frames. Half of the filtered image was combined with half of the delta image to form a 300-dimensional input vector.

In addition to the image basis obtained with local PCA and ICA, we also tested a basis of 300 images each of which contained a gaussian bump padded with zeros. As with local ICA and PCA, the location of the gaussian was different for each of the 300 images in the basis. The size and variance of this bump was adjusted to optimize performance. Figure 4 shows various properties of three local kernels — the gaussian, the first eigenpatch (PCA), and the smallest magnitude ICA filter. All had low-pass properties in the frequency domain, though the local ICA filter had a broader spectrum. The variance images were obtained by computing the variance of the filter output for each of the 300 locations across the whole dataset of 934 images. The local PCA filter had the highest variance filter outputs, followed by the gaussian, and then the ICA filter.

2.3 RECOGNITION ENGINE

The different visual representations described above (global PCA, global ICA, 5 local PCA, 5 local ICA, gaussian) formed the input to hidden Markov models (HMMs) which were separately trained for each word category. The images were modeled as mixtures of gaussian distributions in pixel space. Nine different HMM architectures were tested for each visual representation: 5, 7, and 9-state HMMs with mixture models of 3, 5, or 7 gaussians to represent each state. The initial state probabilities, transition probabilities, mixture coefficients, and mixture centroids were optimized using the EM algorithm. For numerical stability we constrained the variance parameters for aii the states and mixtures to be equal, a constraint that has worked well in previous investigations. The centroids of the mixtures were initialized using a linear segmentation followed by k-means clustering.

3 RESULTS

Generalization performance was computed based on a jackknife training procedure: HMMs were trained on 11 subjects and tested on the remaining 12th subject. Figure 5 shows the test set percent correct for the gaussian filter, 5 local PCA filters, and 5 local ICA filters. The local ICA filter with the smallest magnitude gave the best performance (86.5%). Performance of the remaining ICA filters dropped monotonically. Similarly, the first eigenpatch gave the best local PCA performance. For the remaining PCA filters, performance dropped, but not monotonically. Table 1 shows the best generalization performance for each of the 5 visual representations.

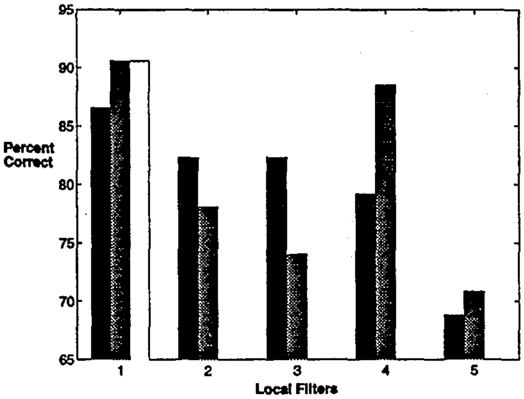

Figure 5: Generalization performance for local filters. Black bars: ICA filters. Gray bars: PCA filters. White bar: Gaussian filter.

Image Processing	Performance
Gaussian	90.6 ± 3.8
Local PCA	90.6 ± 3.1
Local ICA	86.5 ± 4.2
Global PCA	76.0 ± 4.5
Global ICA	78.1 ± 5.4

Table 1: Best generalization performance (% correct) ± s.e.m. for the different visual input representations.

4 DISCUSSION

The results presented here are unambiguous: for the visual speechreading task, image decompositions with local basis images outperform decompositions with global basis images. Even a simple decomposition with local Gaussian bumps significantly outperforms global decompositions obtained with PCA or ICA. These results are consistent with well known properties of visual processing in primary visual cortex, where the existence of local receptive fields suggests a local image decomposition. However, it is unclear why local decompositions work better. One possibility is that these results apply only to this particular recognition engine and the problem at hand (i.e., hidden Markov models for speechreading). Similar results were also obtained by Padgett and Cottrell [7] on a facial expression classification task using a neural network architecture. This supports the idea that there is something fundamentally important about local image decompositions. Another possible explanation is that global decompositions emphasize token information (e.g., subject identity) as opposed to class information (e.g., the word being said). We plan to test this hypothesis in the near future by repeating our analysis with a single user database. Another possibility is that the prior we are placing on our image decompositions (constraining them to be local) is well suited for generalization on the small database we are testing. We are currently extending our database by a factor of 20 to test this hypothesis. Finally, it is possible that there is some fundamental property in the statistics of natural images that makes local decompositions particularly useful for classification tasks. Investigation of this hypothesis may help develop a statistical theory of natural computation that unifies our understanding of the brain and the development of artifical systems designed to process natural signals.

Acknowledgements

We thank Marian Stewart Bartlett, Anthony J. Bell, Martin McKeown, Tzyy-Ping Jung, and Te-Won Lee for helpful discussions. Michael S. Gray was supported by the McDonnell-Pew Center for Cognitive Neuroscience in San Diego, and the Lawrence Livermore National Laboratory ISCR agreement B291528.

References

[1] M.S. Bartlett and T.J. Sejnowski. Viewpoint invariant face recognition using independent component analysis and attractor networks. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, pages 817–823. MIT Press, Cambridge, MA, 1997.

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

[3] A.J. Bell and T.J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, In press.

[4] J. Luettin, N.A. Thacker, and S.W. Beet. Visual speech recognition using active shape models and hidden markov models. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 817–820, Atlanta, Ga, 1996. IEEE.

[5] J.R. Movellan. Visual speech recognition with stochastic networks. In G. Tesauro, D.S. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, pages 851–858. MIT Press, Cambridge, MA, 1995.

[6] B. Olshausen and D.J. Field. Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, 7:333–339, 1996.

[7] C. Padgett and G. Cottrell. Identifying emotion in static face images. In *Proceedings of the Second Joint Symposium on Neural Computation*, volume 5, pages 91–104. Institute for Neural Computation, University of California-San Diego, 1995.

[8] P.S. Penev and J.J. Atick. Local feature analysis: A general statistical theory for object representation. *Network: Computation in Neural Systems*, 7(3):477–500, 1996.

[9] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3:71–86, 1991.