UCSD (1998)

Local Filter Selection Boosts Performance of Automatic Speechreading

Michael S. Gray, Terrence J. Sejnowski
Computational Neurobiology Laboratory
The Salk Institute
La Jolla, CA 92037
michael,terry@salk.edu

Javier R. Movellan
Department of Cognitive Science
University of California, San Diego
La Jolla, CA 92093
movellan@cogsci.ucsd.edu

Abstract

We examine general purpose unsupervised techniques for visual preprocesing in machine vision tasks. In particular we analyze a wide variety of principal component and independent component techniques in combination with stepwise regression methods for variable selection. The task at hand is recognition of the first four digits spoken in English using hidden Markov models (HMM) for the recognition system. Local representations consistently outperformed global representations in generalizing to new speakers while global representations performed better than local ones for speaker identification tasks. In addition, the use of a novel regression-based variable selection technique substantially boosted performance.

1 Introduction

Supervised recognition systems depend on input representations from which class-dependent structure can be easily extracted. In this paper, we explore unsupervised data-driven statistical techniques to develop such image representations and to automatically select variables of interest from high-dimensional outputs. For concreteness we concentrate on the problem of visual speechreading, but the methods explored are general and can be applied to a variety of problems involving recognition of visual sequences. We compare representations obtained with principal component analysis (PCA), independent component analysis (ICA; Bell & Sejnowski [2]) and stepwise multiple regression (Walpole, Myers, & Myers [9], p. 438). In addition, we explore the differences between local and global image representations, a topic of recent interest in the face processing community (Padgett & Cottrell [8]) and in computational neuroscience (Field [4]; Bell & Sejnowski [3]).

The techniques used in this paper attempt to describe efficiently the probabilistic structure in image databases. Such structure can be approached from the point of view of two different probability spaces, which we call image space $(\Omega_{im}, \mathcal{F}_{\rm im}, P_{\rm im})$ and pixel space $(\Omega_{\rm pix}, \mathcal{F}_{\rm pix}, P_{\rm pix})$. Consider an $n \times m$ matrix x whose rows contain the images in the database: $x_{i,j}$ is the intensity of pixel j in image i. In image space, $\Omega_{\rm im} = \{1, \cdots, n\}$, each element of Ω_{im} is given equal probability, and $F_{\rm im}$ is the power set of Ω_{im}. In this space, pixel intensities act as random variables (i.e, functions $\Omega_{\rm im} \to \Re$), and it makes sense to talk about independence between pixels. For example, pixels X_i and X_j are independent if knowledge of the value of pixel X_i in one image does not help estimate the value of pixel X_j in the same image. In pixel space, the situation reverses. The elementary outcomes label the pixels: $\Omega_{\rm pix} = \{1, \cdots, m\}$, and images act as random variables (i.e., functions $\Omega_{\rm pix} \to \Re$). In pixel probability space, two images Y_i and Y_j are independent if knowledge about the intensity of a pixel in image Y_i does not help estimate the intensity of the equivalent pixel in image Y_j. Hereafter we represent pixel intensity with the image-space random vector $X = (X_1, \cdots, X_m)^{\rm T}$, where $X_i(j) = x_{j,i}$.

2 Global Methods

We evaluated unsupervised techniques that operate on whole images as opposed to portions of images. In particular we compared the performace of principal component analysis (PCA) and two different versions of independent component analysis (ICA). We worked with the Tulips1 database (Movellan [7]): 96 digitized movies of 12 undergraduate students (9 males, 3 females) from the Cognitive Science Department at UC-San Diego. The database was normalized by tracking the outlines of the lips using point distribution models (Luettin [5]). Based on the tracked contours, the lip images were normalized for translation and rotation. Finally, the lip images were symmetrized horizontally with respect to the central vertical axis of the image. The images were cropped to 65 pixels vertically × 87 pixels horizontally (5655 pixels total).

Global PCA in Image Space Let $T = \mathbf{e}^{\rm T} X$ represent the principal components of X, i.e., the columns of $\mathbf{e}$ are the eigenvectors of the covariance of X. The principal components are uncorrelated and the eigenvectors (eigenimages) with largest eigenvalues are an efficient set of orthogonal basis images. In our database the first 50 principal components accounted for 94.6% of the variance in the data (trace of $\mathrm{Cov}(X)$). Figure 1 shows the first 5 eigenimages, and their magnitude spectrum. As observed by previous researchers, the basis images obtained via PCA are typically non-local in the spatial domain (i.e., have non-zero energy distributed over the whole image). The principal components T were fed to the HMM recognition engine as described in Section 4.

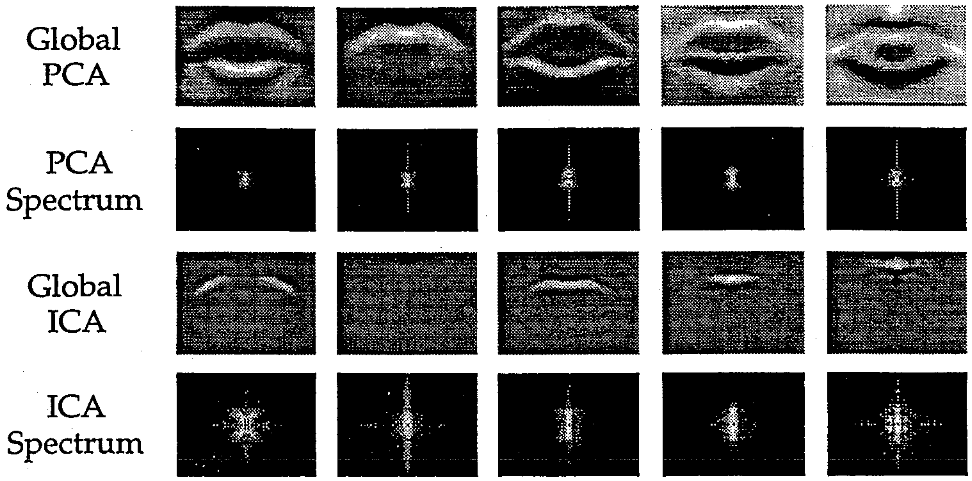

Figure 1: Global decompositions for the normalized dataset. Row 1: Global kernels of principal component analysis ordered with first eigenimage on left. Row 2: Log magnitude spectrum of eigenimages. Row 3: Global pixel space independent component kernels ordered according to projected variance. Row 4: Log magnitude spectrum of global independent components.

Global ICA in Pixel Space The main differences between ICA and PCA are: (1) ICA maximizes the entropy of outputs, while PCA maximizes their variance, (2) PCA provides orthonormal basis vectors, while ICA basis vectors need not be orthogonal, and (3) PCA guarantees uncorrelated components while ICA aims to make the components independent. For computational tractability we applied ICA to the vector of principal components. This results in a vector of independent components $U = \mathbf{w}T$, where $\mathbf{w}$ is a 50×50 matrix found by the ICA algorithm to maximize the joint entropy of the logistic transform of U. The independent components U were fed to the recognition engine.

Global ICA in Pixel Space The procedure described in the previous section maximized joint entropy with respect to the image probability space. An alternative ICA method maximizes entropy with respect to the pixel probability space. This approach has been explored for face recognition tasks (Bartlett et al [1]) and for the analysis of functional magnetic resonance imaging (fMRI) data (McKeown et al [6]) with good results. The approach works as follows: We define a 50-dimensional random vector E in pixel space whose values are the eigenvectors of $\mathrm{Cov}(X)$. We then define the random vector $V = \mathbf{s}E$, where $\mathbf{s}$ is a 50×50 matrix chosen by the ICA algorithm to maximize the joint entropy, in pixel space, of the logistic transform of V.

McKeown et al [6] propose using the independent components in pixel space as basis images in image space. To do so, construct a matrix $\mathbf{v}$ whose columns are the pixel-space independent components (i.e., $\mathbf{v}_{i,j} = V_i(j)$). The goal is to obtain the coordinates of X with respect to the basis formed by the columns of $\mathbf{v}$, and approximate the coordinates of X using the coordinates of $X_{\text{rec}} = \mathbf{e}T$, the reconstruction of X based on the first 50 principal components in image space. It follows that $X \approx X_{\text{rec}} = \mathbf{e}T = \mathbf{v}W$ where W are the desired coordinates of X_{rec} with respect to v. It can be shown that the previous equation is solved for $W_j(i) = (\mathbf{a}\mathbf{s}^{-1})_{i,j}$, where $\mathbf{a}_{i,j} = T_j(i)$.

The coordinates W were the input to the HMM recognition engine.

The first 5 columns of $\mathbf{v}$ (accounting for the largest amounts of projected variance) obtained via ICA analysis in pixel space are shown in the third row of Figure 1. The fourth row shows their magnitude spectrum.

3 Local Decompositions

Recent research has placed strong emphasis on the importance of recognizing *local* structure in images. Analysis of natural images by Field [4] and Bell and Sejnowski [3] has suggested that they can be efficiently represented by spatially localized basis images at a variety of scales. To test the idea that local basis images may be better, we tested a variety of kernels that were spatially *local* (i.e., had non-zero energy only in a small region of the image). Local PCA and ICA kernels were developed based on the statistics of local image regions. Small image patches (12 pixel × 12 pixel) were chosen from random locations in the lip images (similar to Padgett & Cottrell [8]). Twenty patches were randomly 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 top panel of Figure 2. This dataset of local patches (144 pixels × 18680 patches) formed the input to PCA and ICA. Hereafter we refer to the 12 pixel × 12 pixel images obtained via PCA or ICA as "local kernels". For each kernel, basis images were generated by centering a local PCA or ICA kernel onto a particular location of a 65 × 87 matrix and padding the rest of the matrix with zeros, as displayed in Figure 2 (lower left panel).

The efficacy of the local PCA and ICA kernels for recognition was explored using three different approaches: a single filter with linear shift invariant (LSI) filtering, and a bank of filters using blocked or unblocked variable selection.

Single LSI Filtering Images were convolved with a *single* local ICA kernel, local PCA kernel or a Gaussian kernel. This effectively implemented linear shift invariant (LSI) filters. The top 5 local PCA and ICA kernels were each tested separately. We also tested 4 Gaussian kernels of different size. The outputs of these 14 filters were subsampled and *independently* fed to the recognition engine. We report below the performance of the best local PCA filter, the best local ICA filter and the best Gaussian filter.

Bank of LSI Filters with Stepwise Selection Multiple filter representations were used to explore the possibility that combining the outputs of several filters would improve generalization performance. Filter outputs from the top 10 local ICA (or PCA) kernels resulted in a 1500 dimensional representation (10 filters × 150 locations) for each of the 934 images in the dataset. Due to the large dimensionality of the output, we used a stepwise multiple regression procedure to select variables and locations of interest (Walpole, Myers, & Myers [9], p. 438). This method automatically selected those variables that were most informative for reconstruction of the original images. At the first iteration, we constructed 1500 linear regression models, one for each of the 1500 variables. The models were evaluated in terms of their ability to reconstruct the original

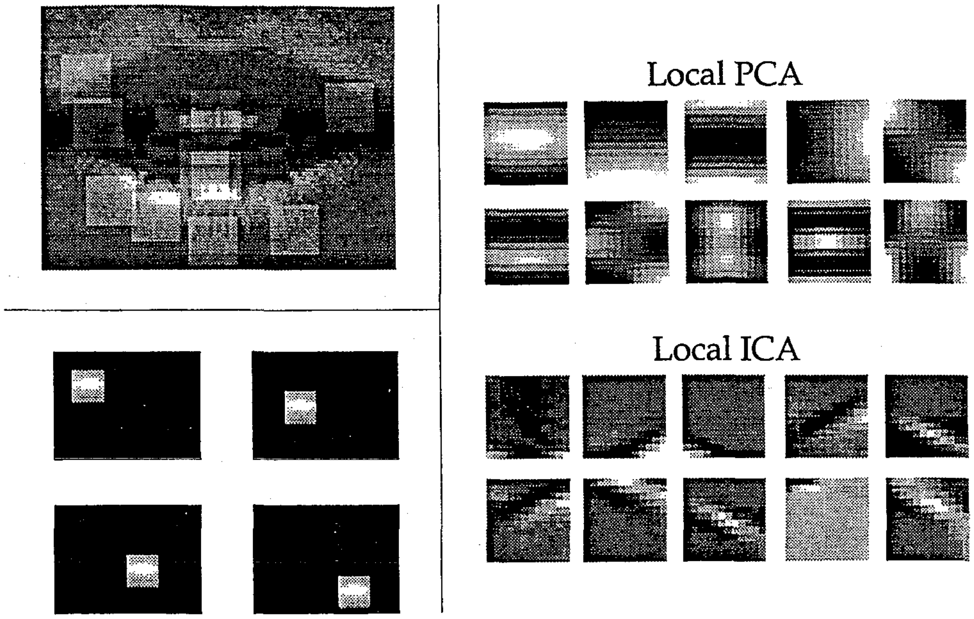

Figure 2: Upper left: Lip patches (12 pixels × 12 pixels) from randomly chosen locations used to develop local PCA and local ICA kernels. Lower left: Four orthogonal basis images generated from a single local PCA kernel. Right: Top 10 Local PCA and ICA kernels ordered according to projected variance (highest at top left).

image dataset. The variable that proved best for reconstruction was "tenured". In subsequent iterations we constructed $1500 - t$ different multiple regression models. Each model contained the t tenured variables plus a non-tenured variable. The variable that provided the best image reconstruction in coordination with the already tenured variables was tenured. The process was stopped when the desired number of tenured variables was reached.

Variable Selection Blocked by Location In this method (Blocked Filter Bank), the images were passed through a bank of 10 LSI filters where the impulse response of each filter corresponded to one of the local PCA or local ICA kernels (Figure 2). After subsampling, this resulted in a 1500 dimensional representation. The stepwise forward multiple regression procedure (described in the previous section) was then used to identify regions of interest. The selection was done in blocks of 10 variables where each block contained the outputs of the 10 filters at a specific location. If a location was chosen, the outputs of the 10 filters in that location were automatically included in the final image representation. Thus the number of outputs per location was either 0 or 10.

Unblocked Variable Selection In this method (Unblocked Filter Bank), the images were passed through the same bank of 10 LSI filters as in the previous approach. However, the forward selection procedure was used without blocking variables by location. Thus the number of selected ouputs per location could vary from 0 to 10.

	Image Processing	Performance ± s.e.m.
Global Methods	Global PCA	79.2 ± 4.7
	Global ICA Image Space	61.5 ± 4.5
	Global ICA Pixel Space	74.0 ± 5.4
Local Methods	Single-Filter LSI PCA	90.6 ± 3.1
	Single-Filter LSI ICA	89.6 ± 3.0
	Single-Filter LSI Gaussian	90.6 ± 3.8
	Blocked Filter Bank PCA	85.4 ± 3.7
	Blocked Filter Bank ICA	85.4 ± 3.0
	Unblocked Filter Bank PCA	91.7 ± 2.8
	Unblocked Filter Bank ICA	91.7 ± 3.2

Table 1: Best generalization performance (% correct) ± standard error of the mean for all image representations.

4 Results

The image representations obtained using each of the processing methods were fed to a recognition engine. This engine first computed delta vectors (the difference between temporally adjacent input vectors) and scaled the vectors using an adaptive thresholding procedure. The scaled vectors were fed to a bank of HMMs consisting of 4 HMMs separately trained on each digit. The architecture was left-right with state skips allowed. The density model for the observations was a mixture of Gaussian distributions. 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. Generalization performance for each visual representation was computed based on the jack-knife procedure. This was repeated 12 times, each leaving out a different subject. This procedure also allows obtaining classical confidence intervals on the generalization score. Table 1 shows the best generalization performance (of the 9 HMM architectures tested) for all visual representations tested. The local decompositions significantly outperformed the global representations: $t(106) = 4.10$, $p < 0.001$. In addition, for the filter bank representations, the unblocked approach yielded better results than the blocked: $t(46) = 1.95$, $p < 0.06$.

The image representations obtained using the bank of filter methods with unblocked selection yielded the best recognition results. Figure 3 shows, for 2 local PCA kernels, the first 10 variables chosen for each particular kernel using the forward selection multiple regression procedure. The numbers on the lip images in this figure indicate the order in which particular kernel/location variables were chosen using the sequential regression procedure: "1" indicates the first variable chosen, "2" the second, etc. In total, there were 50, 100, or 150 kernel-location variables chosen for the PCA and ICA representations (see Section 3).

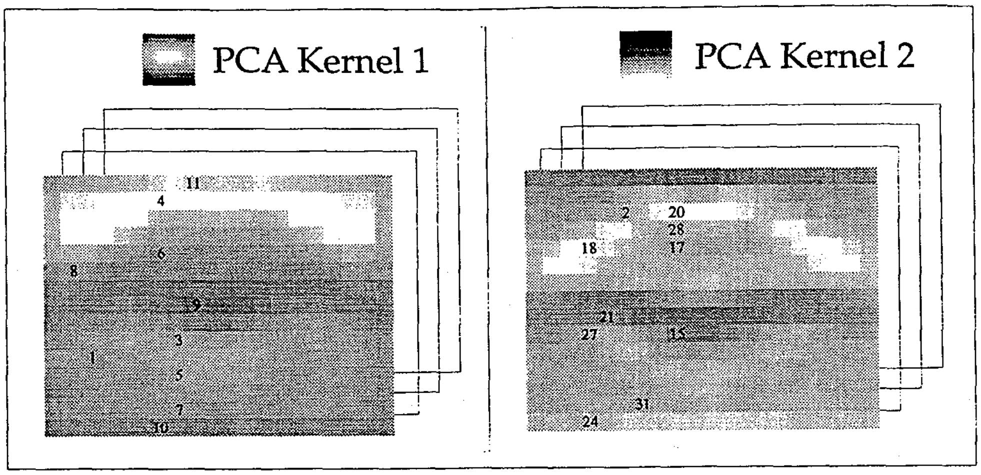

Figure 3: Kernel-location combinations chosen using unblocked variable selection. Top of each quadrant: Local ICA or PCA kernel. Bottom of each quadrant: Lip image convolved with corresponding local kernel, then downsampled. The numbers on the lip image indicate selected variables. There are no numbers on the right side of the lip images because only half of each lip image was used for the representation.

5 Discussion

The experiments described here yielded two primary findings. First, unsupervised statistical image decompositions with local basis images outperformed decompositions with global basis images. The highest generalization performance reported here (91.7% with the bank of filters using unblocked variable selection) surpasses the best published performance on this dataset (Luettin [5]). Even a simple decomposition with local Gaussian kernels significantly outperformed global decompositions obtained via PCA or ICA. Second, the stepwise regression technique used to select variables and regions of interest led to substantial gains in recognition performance. Figure 3 shows the first 8-10 points chosen from the local PCA and ICA kernel outputs. The chosen locations (variables) roughly followed the contour of the lips.

The superior performance of local representations is consistent with current ideas on the importance of locality (see Section 3). One possible explanation for the advantage of local representations (Padgett & Cottrell [8]) is that global unsupervised decompositions emphasize subject identity since it is an important source of variation. In speaker independent tasks (e.g. recognizing the word being said), subject identity is precisely what needs to be deemphasized. We tested this idea with a simple subject recognition task on the Tulips1 database. The task was to recognize the identity of the speaker in each of the 934 images in the database. The recognition engine was a simple prototype classifier (a bank of HMMs, one per subject to be identified, with a single state and a single Gaussian for the observation density). We compared subject identification performance using our best global representation (Global PCA) and our best local representations (Unblocked Filter Bank ICA and PCA). The difference in performance between the local and global representations was astounding. For the local representations, subject identity was recovered with 39.0% accuracy

(ICA) and 44.5% accuracy (PCA). Global representations recovered subject identity with a 94.8% accuracy. Thes results suggest that local representations are better for speaker-independent tasks and that holistic representations may be more appropriate for speaker identification problems.

References

[1] M. S. Bartlett, H. M. Lades, and T.J. Sejnowski. Independent component representations for face recognition. In *Proceedings of the SPIE Symposium on Electronic Imaging: Science and Technology*, volume 3299, San Jose, CA, In press. SPIE Press.

[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*, 37(23):3327–3338, 1997.

[4] D.J. Field. What is the goal of sensory coding? *Neural Computation*, 6:559–601, 1994.

[5] Juergen Luettin. *Visual Speech and Speaker Recognition.* PhD thesis, University of Sheffield, 1997.

[6] M.J. McKeown, T.-P. Jung, S. Makeig, G. Brown, S.S. Kindermann, T.-W. Lee, and T.J. Sejnowski. Spatially independent activity patterns in functional MRI data during the stroop color-naming task. *Proceedings of the National Academy of Sciences*, 95:803–810, 1998.

[7] 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.

[8] C. Padgett and G. Cottrell. Representing face images for emotion classification. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, pages 894–900. MIT Press, Cambridge, MA, 1997.

[9] D. Walpole, R.H. Myers, and S.L. Myers. *Probability and Statistics for Engineers and Scientists.* Prentice Hall, Upper Saddle River, NJ, 1998.