Sparse PCA
Extracting Multi-scale Structure from Data

Chakra Chennubhotla & Allan Jepson
Department of Computer Science
University of Toronto, 6 King's College Road
Toronto, ON M5S 3H5, Canada
Email: {chakra,jepson}@cs.toronto.edu

Abstract

Sparse Principal Component Analysis (S-PCA) is a novel framework for learning a linear, orthonormal basis representation for structure intrinsic to an ensemble of images. S-PCA is based on the discovery that natural images exhibit structure in a low-dimensional subspace in a sparse, scale-dependent form. The S-PCA basis optimizes an objective function which trades off correlations among output coefficients for sparsity in the description of basis vector elements. This objective function is minimized by a simple, robust and highly scalable adaptation algorithm, consisting of successive planar rotations of pairs of basis vectors. The formulation of S-PCA is novel in that multi-scale representations emerge for a variety of ensembles including face images, images from outdoor scenes and a database of optical flow vectors representing a motion class.

1 Introduction

Our goal is to extract structure intrinsic to an ensemble of images. We consider ensembles formed from natural images—for example, images sampled from outdoor environments, face images acquired from roughly similar viewpoints, images of a gesturing hand, or vector-valued optical flow images collected from motion sequences. As is well known, the images in such ensembles exhibit structure at multiple spatial scales. In this paper we describe a new learning technique, for deriving a linear basis representation, to highlight ensemble-specific, multi-scale structure.

To obtain a multi-scale representation, there are at least two approaches to take: (1) use a basis set that is "predefined" or "fixed", e.g. 2-D Gabors or wavelets; (2) *learn* a representation to match the structure in an image ensemble. Since a predefined basis is inflexible, and often awkward to define, we pursue a learning framework instead.

Most learning algorithms are based on the hypothesis that images are caused by a linear combination of statistically *independent* components. The typical goal is to seek a representation that can reduce, if not eliminate, pixel redundancies. The hope is that basis functions will acquire the shape and form of the underlying independent component structure. One strategy is to combine the knowledge of ensemble statistics with simple optimization principles. For example, *Principal Component Analysis* (PCA) [7] uses second-order statistics to decorrelate the outputs of an orthogonal set of basis vectors. Alternatively, *Sparse coding* constrains outputs to be drawn from a low-entropy distribution to achieve, if not independence, at least a reduction in higher-order dependencies [12]. Similarly, *Independent Component Analysis* (ICA) is closely related to sparse coding and is based on an information-theoretic argument of maximizing the joint entropy of a non-linear transform of the coefficient vectors [2].

The strategies of sparse coding and ICA were deemed successful when applied to an ensemble of natural scenes, because they extract multi-scale, wavelet-like structure. However, when the input ensemble is specific to an object (e.g. face images), sparse coding, ICA and PCA lead to basis images that are *not* multi-scale, appear *holistic* and lack an "obvious" visual interpretation [1, 15, 17]. We will explore this observation further when we compare these methods with our framework below.

A primary contribution of this paper is to show how multi-scale representations emerge, for a variety of image ensembles, by trading off *redundancy minimization* for *sparsity maximization* in a basis matrix. We base our strategy, *Sparse Principal Component Analysis* (S-PCA), on the discovery that natural images exhibit structure in a *low-dimensional* subspace in a *sparse*, *scale*-dependent form. As we demonstrate, placing a sparse prior on the basis elements is a powerful way to predispose the learning mechanism to converge to this naturally-occurring, sparse, multi-scale structure. We will show that, while PCA determines the optimal subspace, S-PCA discovers the structure internal to that space.

S-PCA learns an orthonormal basis by rotating the basis vectors that span the principal subspace. Rotation achieves sparsity in the basis vectors at the cost of introducing correlations in the output coefficients. If the input ensemble is a multi-dimensional Gaussian with widely separated variance distribution, then S-PCA returns a redundancy minimizing solution, namely the basis vectors of PCA. On the other hand, if the input ensemble is devoid of any structure (i.e. i.i.d. pixel intensities), then S-PCA returns a maximally

0-7695-1143-0/01 $10.00 © 2001 IEEE

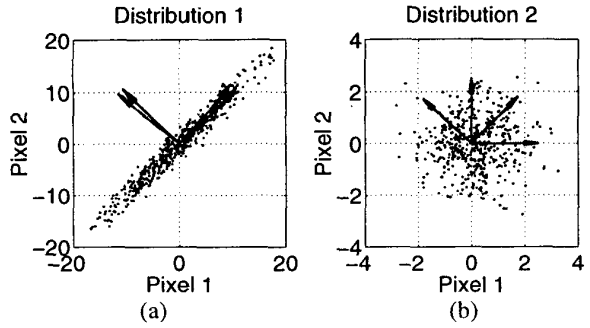


Figure 1: What is PCA good for? Distributions 1 and 2 (black dots) are 2-pixel image ensembles sampled from multi-dimensional Gaussian priors. (a) Distribution 1 has a dominant orientation indicated by the PCA basis (blue). (b) Distribution 2 has no orientational structure, and the PCA basis (blue) reflects sampling noise. In both cases the preferred S-PCA basis vectors are obtained by rotating PCA directions. In (a) the rotation is minimal, while in (b) the rotation maximizes sparsity in the basis vector description by aligning them with the pixel basis.

sparse representation, with each basis vector representing the brightness at a single pixel. As our examples show, this property of returning a sparse representation, when possible, provides a much more intuitive representation of the ensemble than a standard PCA based representation.

Besides this theoretical advantage of providing a better intuitive model of an ensemble, the computation of the basis coefficients is more efficient because of the presence of zero valued weights in the sparse basis vectors. The speed-up obtained from this sparsity will depend on both the ensemble and on the number of basis vectors used. In particular, for all the natural data sets we have considered, the S-PCA basis vectors tend to get increasingly sparse as the corresponding variances decrease. Therefore the speed-up due to sparsity can be very significant when many basis vectors are used, but less significant when only a few are used.

2 Encouraging Sparsity

To illustrate the basic idea of S-PCA, consider 2-pixel image ensembles generated from a multi-dimensional Gaussian distribution. Each image is a point in a 2-dimensional space. In Fig. 1, we show two such datasets, one with a dominant orientation (a) and the other essentially isotropic (b).

PCA determines an orthogonal set of basis vectors with the property that the basis expansion coefficients are decorrelated. The idea of PCA in this 2D example is to rotate the pixel basis, $\binom{1\,0}{0\,1}$, until the variance of the projected data is maximized for one component and is minimized for the other. In Fig. 1, the PCA directions for the distributions are shown in blue. For the correlated dataset in Fig. 1a, the PCA vectors are aligned with the oriented structure underneath. For the uncorrelated dataset in Fig. 1b, the specific directions that PCA selects are dictated by sampling noise. For such datasets we prefer basis vectors that are sparse, that is, have few non-zero entries. In this 2D example, this is just the pixel basis, $\binom{1\,0}{0\,1}$, as shown in red in Fig. 1b. Note the sparse basis can be achieved by simply *rotating* the PCA basis by an amount depending on the degree of correlation in the underlying dataset.

3 S-PCA Pairwise Rotation Algorithm

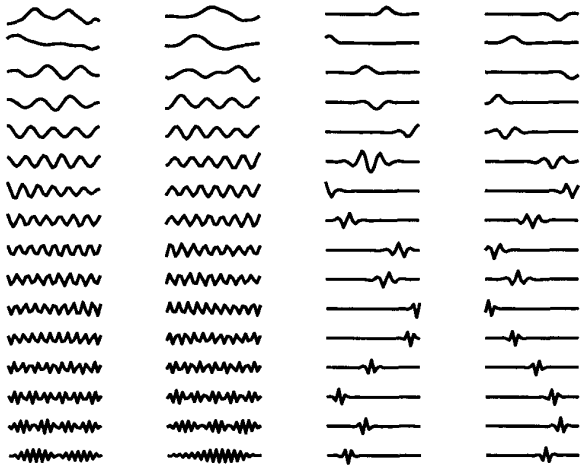
The idea behind S-PCA is to retain the PCA directions when there is correlational structure in the data set, and otherwise rotate them to be as sparse as possible. We propose a cost function $C(\lambda) = C_1 + \lambda C_2$, where C_1 is a function of the variances of the data projected onto the individual basis vectors, and C_2 is a function of the elements of the basis vectors themselves. The exact form for C_1 and C_2 are not so important, so long as C_1 is designed to retain the PCA directions while C_2 promotes sparsity. The λ parameter in the cost function provides the relative importance of the sparsity term, and in this paper we choose it to make the contributions from C_1 and C_2 have the same scale. See the Appendix for the actual C_1, C_2 and λ used in all the examples in this paper.

The learning algorithm of S-PCA is very simple. The basis vectors are initialized to be the principal components. The dimensionality of this principal subspace is chosen before hand. Every pair of these basis vectors defines a hyperplane, and we successively select suitable rotations within these hyperplanes to minimize $C(\lambda)$. We sweep through every possible basis vector pair doing these rotations, and these sweeps are repeated until the change in $C(\lambda)$ is below a threshold. The product of the pairwise rotations provides a composite rotation matrix which, when applied to the PCA vectors generates the S-PCA basis. In particular, the S-PCA and PCA bases are orthonormal representations for the same subspace and are related to each by other by this rotation matrix. A typical implementation of this algorithm is included in the Appendix.

The PCA basis is used as the starting point since it identifies the principal subspace best suited to recovering correlational structure. The job of the S-PCA algorithm is then simply to resolve the range of the spatial correlations. Note that the S-PCA basis is always a rotation of the original basis, so some care should be taken in choosing this starting basis. In cases for which we want a complete representation, we have found that the trivial basis (i.e. provided by the columns of the identity matrix) can also be used as a starting point for the S-PCA algorithm.

3.1 S-PCA on Filtered Noise

For an illustrative example we consider an ensemble generated by convolving random vectors $\vec{x}$ with a Gaussian kernel $\vec{g}$, and adding small amounts of noise $\vec{n}$. That is,



(a) PCA (b) S-PCA

Figure 2: Representing filtered noise ensemble. Each waveform is 32 pixels long. (a) Basis vectors from PCA (b) Multi-scale basis vectors from S-PCA. Zero valued weights cause basis matrix to be sparse.

$\vec{y} = \vec{g} \otimes \vec{x} + \vec{n}$, with both $\vec{x}$ and $\vec{n}$ i.i.d. signals. Note that, because of the smoothing, the output signal $\vec{y}$ is correlated in spatially local neighborhoods. Below we show that S-PCA provides a representation which highlights this structure.

A Principal Component Analysis (PCA) of such a data set provides a basis representation which is global and spatial frequency specific. The PCA basis is depicted in Fig. 2a, where each waveform corresponds to a basis vector. Notice that this representation does not highlight the spatially localized structure introduced by the Gaussian smoothing.

The S-PCA basis derived for the low-pass filtered noise vectors provides structure at multiple scales (see Fig. 2b). The first few basis vectors appear as low-pass signals, whose size is indicative of the Gaussian kernel used to introduce the correlations between pixels. While these basis vectors are spread evenly over the pixel space, they cannot significantly overlap due to the orthogonality constraint. Instead, basis vectors at the next scale exhibit entries with multiple signs, but still remain local. Thus S-PCA generates spatially localized, bandpass functions as basis vectors, thereby achieving a joint space and frequency description in a manner similar to a wavelet basis. Moreover, the basis derived by S-PCA is orthogonal.

The separation in scales can be also seen in the plot of the variances as a function of the basis index (Fig. 3a). Notice that the smoothly decaying variances for the the PCA basis have been reshaped by S-PCA to locally flat portions. Each flat portion corresponds to basis functions at a particular scale.

There are two other properties of S-PCA worth noting. First, when λ is non-zero, the fraction of input variance captured by the first k-dimensional subspace is always higher in PCA (blue curve in Fig. 3b) than S-PCA (red curve in Fig. 3b). However, the difference between these two curves is relatively small, compared to the increased sparsity of the basis (see Fig. 2). As λ is increased, the sparsity of the S-PCA basis is increased, with a corresponding decrease in the variance captured by S-PCA. At extremely large values of λ, the S-PCA basis becomes maximally sparse, with each basis function representing an individual pixel. The variance captured by the maximally sparse basis is given by the green line in Fig. 3b. The second property of S-PCA is that the rotation of PCA basis introduces small correlations in the output coefficients (see Fig. 3c,d).

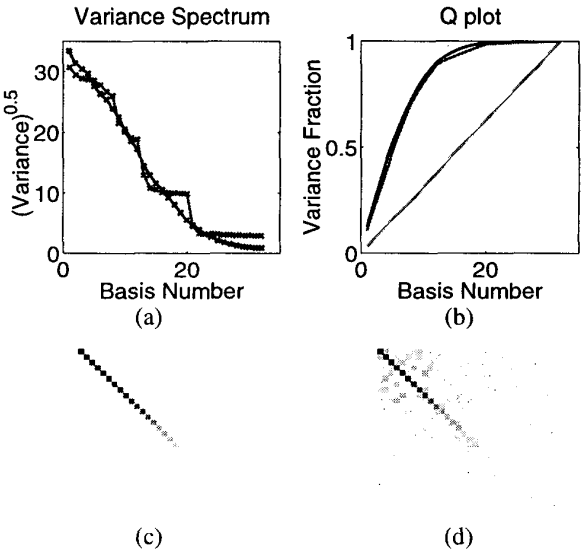


Figure 3: Representing filtered noise ensemble (contd). (a) Smoothly decaying variances of PCA (blue) reshaped to locally flat portions by S-PCA (red) (b) Fraction of the total variance captured by k-dim subspaces as k varies from 1 to 32: PCA (blue), S-PCA (red), Identity matrix (maximally sparse) (green). The covariance matrix is diagonal for PCA (c) but only diagonally dominant for S-PCA (d).

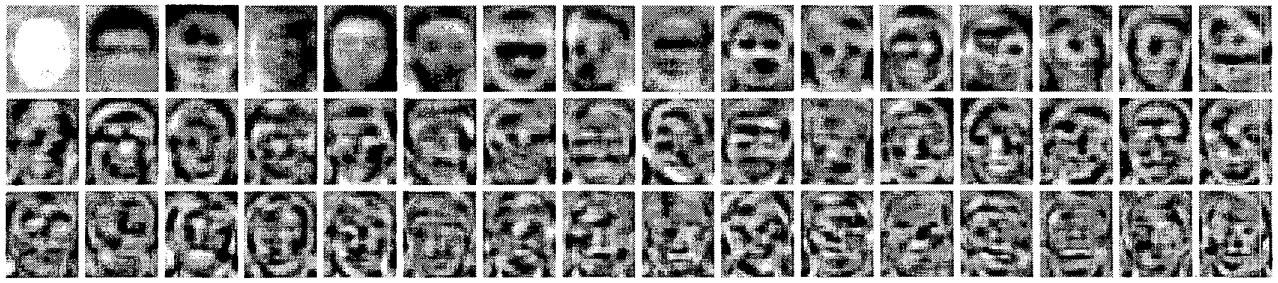
4 S-PCA on Natural Ensembles

We next apply S-PCA on ensembles of natural images. For each ensemble, we present our guess for the inherent structure and show that S-PCA confirms our intuition.

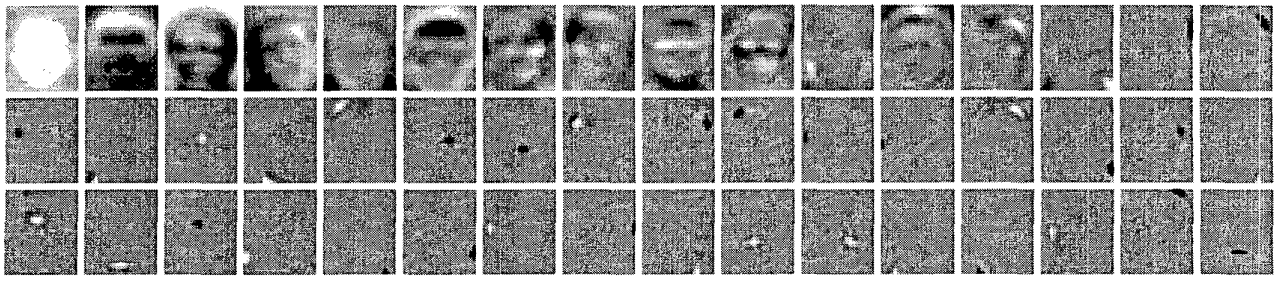
4.1 Facelets

We represent face images that have been acquired under roughly similar viewing conditions with a varying degree of pose changes. For such an image collection we argue that the ensemble can be decomposed using a basis set of spatially coherent image blobs. The spatial coherence is caused by the consistency and uniqueness in the overall shape of the object and in the shape of its parts across the ensemble.

The PCA representation displayed in Fig. 4a is typical



(a) PCA



(b) S-PCA

Figure 4: *Facelets* for facial images each of size: 28×23. The top 200 PCA basis vectors were reorganized to generate S-PCA basis. Only the first 48 basis vectors are displayed here for (a) PCA (b) S-PCA. While PCA basis appear global, S-PCA is multi-scale. Subimages are rescaled to have zero value correspond to a gray level of 127.

of the results obtained for object specific image ensembles. In particular, the first few principal vectors are relatively smooth and represent global correlations. As the index k increases (i.e. the variance of the component decreases), the basis vectors represent finer spatial structure but remain global in space. This is apparent in Fig. 4a from the relative sizes of the individual regions in which the basis functions are of one sign, which get increasingly smaller with increasing index. Unfortunately, it is impossible to tell from the PCA basis whether or not such fine scale structure is correlated across the entire image.

The first few basis vectors for the S-PCA results on the face ensemble represent correlations across the whole image (see Fig. 4b). However, as the index increases, the basis vectors become quite sparse, indicating that the information at finer spatial scales is not strongly correlated across the entire image. Indeed, 72.5% of the elements of the first 40 S-PCA basis vectors are less than 5% of the maximum absolute value, and can be thresholded to zero (see Fig. 5). This proportion of effectively zero elements increases as larger dimensional basis sets are considered.

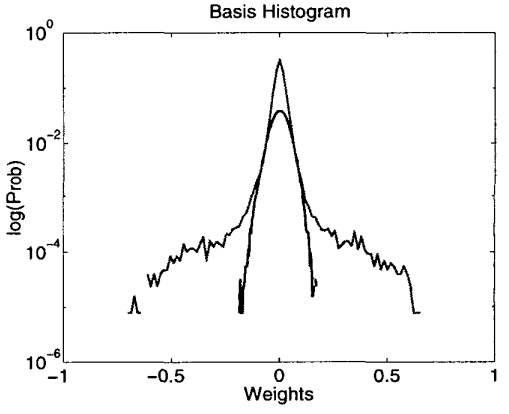


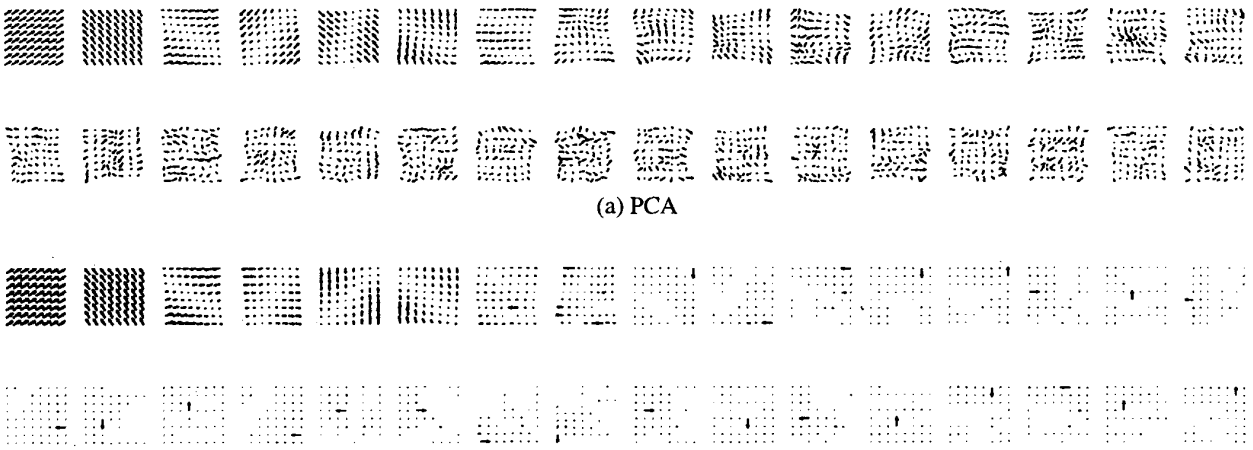
Figure 5: Histogram of the basis elements for PCA (blue) and S-PCA basis (red). Note the increase in probability volume around zero for S-PCA.

4.2 Flowlets

The results of PCA and S-PCA on a set of flow fields obtained for a bush blowing in the wind are presented in Fig. 6. We see that the PCA results provide global vector fields which vary on increasingly fine scales as the basis index increases (Fig. 6a). The PCA analysis leaves open the question of whether or not the fine scale structure of the flow is correlated across the image patch. However, S-PCA reveals that the local flow is predominantly affine (Fig. 6b), with the finer scale structure collapsing down to variations involving essentially individual pixels.

4.3 Wavelets

The results of PCA and S-PCA on a set of 8×8 patches from natural images show the same general results (Fig. 7). We see that the PCA results provide global basis vectors which represent variations on increasingly fine spatial scales



(a) PCA

(b) S-PCA

Figure 6: *Flowlets* represent optical flow vectors measured from an image sequence of a bush blowing in the wind. The flow fields are sampled in 8×8 pixel blocks. The basis flows from S-PCA provide a clear interpretation of the motion being mostly affine (b), which is not so apparent with PCA (a).

as the basis index increases (see Fig. 7a). Again, the PCA analysis leaves open the question of whether or not this fine scale spatial structure is correlated across the image patch. However, S-PCA reveals that the fine grain structure is not significantly correlated over large spatial regions. Rather, as the basis index increases, the S-PCA basis vectors exhibit increasingly local structure (Fig. 7b). Here the fine scale basis vectors appear to have either a center-surround structure or are similar to orientation tuned, scale-specific filter kernels.

5 Application

For a demonstration of an application for the S-PCA representation, consider the Facelets learned in Fig. 4. As mentioned above, 72.5% of the elements in the top 40 vectors of the S-PCA basis are less than 5% of the amplitude of the corresponding basis vector. We would like to take advantage of these nearly zero elements and avoid doing extra arithmetic operations. Indeed, we set these elements to zero and considered the reconstruction using the resulting thresholded S-PCA basis vectors. It is important to remember that thresholded S-PCA basis vectors need not be strictly orthogonal and hence, the algorithms that exploit sparsity have to take this fact into account.

We found that the average squared reconstruction error, taken over the set of training images and using the thresholded S-PCA basis, was 4.126×10^6, as compared to 4.121×10^6 for the unthresholded S-PCA basis. Thus the thresholding to zero of over 70% of the elements of the S-PCA basis had only an minor impact on the reconstruction error, clearly supporting our claim that the basis is sparse.

As we pointed out above, the sparsity in the S-PCA basis is achieved at the cost of some correlation between the basis vectors. The presence of this correlation implies that the least squares reconstruction error of the S-PCA basis will not be optimal (cf. Fig. 3b). The minimum least squares reconstruction error is obtained by the PCA basis, and for 40 basis vectors, we found the average squared error to be 4.100×10^6. This is to be compared to the averaged squared error of 4.126×10^6, for the thresholded S-PCA basis. Thus we have achieved 70% sparsity in this basis with only a 0.6% increase in the squared error. As a note of caution, λ decides the amount of sparsity in the basis matrix and consequently the reconstruction error. It remains to be seen what the best value for λ is, but for this paper we selected basis vectors for each individual image based on the criterion of minimizing its reconstruction error.

6 Related Work

To the best of our knowledge there is no other algorithm that extract multi-scale structure in object-specific ensembles. There are, however, a host of formulations all closely related to sparse coding [5, 12] and ICA [2], which may appear confusingly similar to our framework, S-PCA. We first explore these connections before highlighting other relevant material.

To begin with, all these methods represent images as linear superposition of basis vectors, i.e. $\vec{t} = U\vec{c}$, where $\vec{t}$ is a N-element image from the input ensemble, U is $N \times M$ matrix whose columns are basis vectors and $\vec{c}$ is a M-element coefficient vector. The idea is to infer the basis matrix U given an image ensemble $\{\vec{t}_i\}_{i=1..k}$. It is useful to note that ICA learns a filter matrix F first, always of size $N \times N$, i.e. $F\vec{t} = \vec{c}$, whose inverse then corresponds to a basis matrix, i.e. $U = F^{-1}$. The resulting coefficients from ICA will be N-element long.

The optimization criterion used in sparse coding (and in-

directly in ICA) is to employ a sparse prior on the elements of the coefficient vector $\vec{c}$. The motivation for the sparse prior comes from observing the shape of the output coefficient histograms of a multi-scale wavelet transformed natural scene. For natural scenes sparse coding, as well as ICA, lead to basis vectors that are multi-scale, non-orthogonal. Unfortunately, the sparse coding algorithm is restricted to work on small sub-images extracted from a larger image. As such, spatial structure larger than the image block may never be captured. The object-specific ensembles (face images) we handle are large in size, so it is not easy to train these algorithms. Instead we refer the reader to the ICA results published by Bartlett et al in [1]. Note that ICA produces filters, not basis functions, and while the filters in [1] appear to capture facial features at one scale, the corresponding basis functions do not.

Varimax is one of the earliest known references for rotating principal components so that the basis vectors are more interpretive [8]. The rotation is performed to increase the variance of the elements that describe the basis vectors. An increase in variance can be seen as an increase in entropy, but as we have shown for interpretable basis directions, the basis elements can be drawn from a distribution with high kurtosis.

The idea of applying a sparse constraint on the weights of a neural net appears in [11, 18]. The networks are non-linear and the learning has not shown the ability to extract multi-scale structure. Recently, such sparsity constraints were extended to both the basis matrix and the coefficients as in [6] but this algorithm is likely to have problems with the size of the input image. Our learning algorithm has a flavor similar to the one used in [3] where the tensor properties of the cumulants for the coefficient densities are exploited using a rotation based Jacobi algorithm.

The notion of sparsity also appears in the basis selection literature. The idea is to have a dictionary of basis functions, possibly multi-scale, and pick a small number of them to represent the input image [4]. The basis functions are either predefined, as in wavelets, or specific to a class, as in correlation-based kernels [14, 13]. Unfortunately, determining coefficients for each image is formulated as a quadratic programming problem and this can be computationally very expensive. In [10] a constraint of positiveness on the basis elements lead to facial features at a single scale.

7 Conclusions

We presented a novel framework, Sparse Principal Component Analysis (S-PCA), for learning a linear, orthonormal basis representation for representing structure intrinsic to an ensemble. We showed how introducing a sparsity constraint on the elements of a basis matrix recovers structure in spatially coherent image blobs and provides a multi-scale representation for the ensemble. The principal advantages of S-PCA over a standard PCA based representation include an intuitive understanding of the features underlying the ensemble and efficiency in computations resulting from a sparse basis representation.

Sparsity in the basis matrix is best understood as saving computations over the lifetime of the representation system. In the process, extra bits are spent in representing images as S-PCA causes output coefficients to be slightly correlated. The learning algorithm of S-PCA is very simple, with an optimization procedure that is robust and scalable to large-dimensional spaces. As we have shown, the formulation of S-PCA is novel in that multi-scale representations emerge for a wide variety of ensembles.

There are several avenues open for research. The S-PCA formulation we presented here does not consider noise in the dataset. We have designed a new algorithm, Sparse Information Maximization, to account for both the photon noise at the input and quantization noise at the output [16]. Finally, it is natural to consider the use of the local basis representations provided by S-PCA for representing partly occluded objects.

Acknowledgements: The first author thanks: Alex Vasilescu, Sageev Oore and Jonathan Appavoo for comments at various stages of this work, IBM Center for Advanced Study Toronto and IRIS Canada for financial support.

References

[1] Bartlett, M., H. Lades and T. J. Sejnowski. Independent Component Representations for face recognition *Proc. SPIE Symposium on Electronic Imaging: Human Vision and Electronic Image III*, v 3299, January 1998.

[2] Bell, A. J. and T. J. Sejnowski. The Independent Components of natural scenes are edge filters. *Vision Research*, 37:3327-3338. 1997.

[3] Cardoso, J. High-order contrast for independent component analysis. *Neural Computation*, 11:157-192. 1999.

[4] Chen, S., D. L. Donoho and M. A. Saunders. Atomic Decomposition by Basis Pursuit *SIAM Journal of Sci. Computing*, 20(1):33-61, 1998.

[5] Harpur, G. F. and R. W. Prager. Development of Low-entropy coding in a recurrent framework. *Network: Computation in Neural Systems* , 7(2):277-284, 1996.

[6] Hyvrinen A. and R. Karthikesh. Sparse priors on the mixing matrix in independent component analysis. *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)*, pp. 477-452, Helsinki, Finland, 2000.

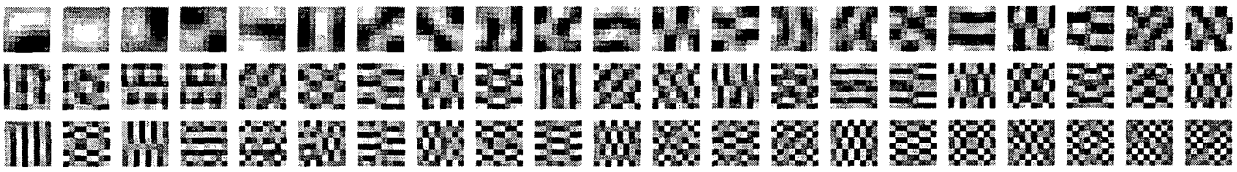
[7] Jolliffe, I. T. Principal Component Analysis. Springer, New York, 1986.

[8] Kaiser, H. F. The Varimax criterion for analytic rotation in factor analysis. Psychometrika 23:187-200. 1958.

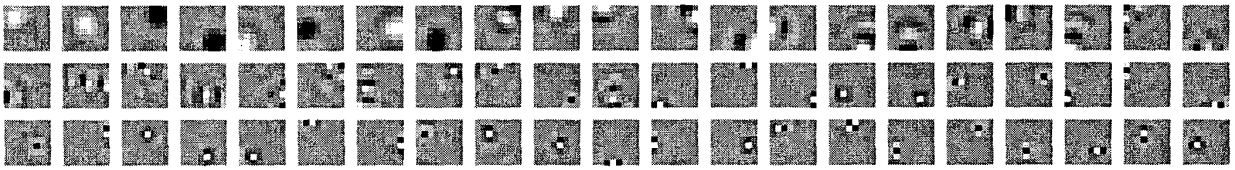
[9] Devijver, P. A and J. Kittler. Pattern Recognition: A Statistical Approach. Prentice Hall, 1982.

[10] Lee, D. and S. Seung. Learning the parts of an object by non-negative matrix factorization. Nature, 401, Oct 1999.

[11] Nowlan S. and Hinton G. Simplifying neural networks by soft weight sharing. *Neural Computation*, 4(4):473:493, 1992.



(a) PCA



(b) S-PCA

Figure 7: *Wavelets* represent images from outdoor scenes. The ensemble is a collection of image patches, each of size 8×8, randomly sampled from natural scenes. (a) PCA basis vectors are global while (b) S-PCA learns a sparse, multi-scale representation.

[12] Olshausen, B. A. and D. J. Field. Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Research*, 37,3311-3325. 1997.

[13] Penev, P. S. and J. J. Atick. Local Feature Analysis. *Network: Computation in Neural Systems*, 7(3), 477-500, 1996.

[14] Papageorgiou C. P., F. Girosi and T. Poggio. Sparse Correlational Kernal Analysis and Reconstruction *ICASSP*, Pheonix, March 1999.

[15] Sirovich, L and M. Kirby. Low-dimensional procedure for the characterization of human faces. *JOSA*, 4, 519-524, 1987.

[16] Chakra Chennubhotla. Sparse Information Maximization, Thesis Manuscript in preparation. 2001.

[17] Turk, M. A. and A. P. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3(1):71-86, 1991.

[18] Williams, P. M. Bayesian regularization and pruning using a Laplace prior *Neural Computation*, 7(1):117-143, 1995.

Appendix

A probabilistic treatment for the cost function, $C = C_1 + \lambda C_2$, is presented in detail in [16]. For this paper, we chose an entropy-like measures for both C_1 and C_2, as given below.

Let $U = [\vec{u}_1 \vec{u}_2 \cdots \vec{u}_M]$ be a basis matrix spanning a $M-$dimensional principal subspace, with $\vec{u_m} = (u_{m,1}, \ldots, u_{m,N})^T$. Let σ_m^2 be the variances of the data projected on the direction $\vec{u}_m$. Set $\vec{d} = (d_1, \ldots, d_M)^T$ to be the vector of relative variances for each of the basis functions, that is, $d_m = \sigma_m^2 / \sum_{k=1}^{M} \sigma_k^2$.

Then the first term of the cost function, namely $C_1(\vec{d})$, is defined to be $C_1(\vec{d}) = \sum_{m=1}^{M} -d_m \log(d_m)$. It can be shown that only when the basis vectors are PCA directions is the cost function C_1 minimized [9].

The second term of the cost function, $C_2(U)$, is defined to be $C_2(U) = \sum_{m=1}^{M} \sum_{n=1}^{N} -u_{m,n}^2 \log(u_{m,n}^2)$. Notice that this is just the sum of the entropies of the distributions defined by the square of the elements for each basis vector $\vec{u}_m$ (recall the basis vectors have unit norm). If the elements of the basis vector have a Gaussian-like distribution, as in PCA, entropy is high and so is the cost function C_2. If the basis vectors from an Identity matrix, entropy is low. Thus, $C_2(U)$ can be seen as promoting sparsity. We chose the λ parameter to make the contributions from C_1 and C_2 have the same scale: $\lambda = (N * \log(M))^{-1}$. A typical implementation of S-PCA is given below:

Algorithm 1 Sparse Principal Component Analysis

```
Define M, N, U, d, lambda as above

done    = 0
oldcost= C_1(d) + lambda*C_2(U)
while (!done)
 for i = 1 to (M-1)
  for j = i+1 to M
   (u_i, u_j) = (i^th, j^th) basis vectors
   (d_i, d_j) = relative variance captured by
                (u_i, u_j)
   rotationAng <- minimize(C_1(d_i, d_j) +
                           lambda*C_2(u_i, u_j)
                          )
   if (rotationAng > thresholdAng)
     update u_i and u_j
     update d_i and d_j
   end
  end
 end
 newcost= C_1(d) + lambda*C_2(U)
 if (abs(oldcost-newcost)/oldcost) < eps
  done = 1
 end
end
```