On Intrinsic Generalization of Low Dimensional Representations of Images for Recognition

Xiuwen Liu*, Anuj Srivastava†, DeLiang Wang‡
*Department of Computer Science, Florida State University, Tallahassee, FL 32306-4530 (Email: liux@cs.fsu.edu)
† Department of Statistics, Florida State University, Tallahassee, FL 32306-4330 (Email: anuj@stat.fsu.edu)
‡ Department of Computer & Information Science and Center for Cognitive Science, The Ohio State University
Columbus, OH 43210 (Email: dwang@cis.ohio-state.edu)

***Abstract*— Low dimensional representations of images impose equivalence relations in the image space; the induced equivalence class of an image is named as its intrinsic generalization. The intrinsic generalization of a representation provides a novel way to measure its generalization and leads to more fundamental insights than the commonly used recognition performance, which is heavily influenced by the choice of training and test data. We demonstrate the limitations of linear subspace representations by sampling their intrinsic generalization, and propose a nonlinear representation that overcomes these limitations. The proposed representation projects images nonlinearly into the marginal densities of their filter responses, followed by linear projections of the marginals. We have used experiments on large datasets to show that the representations that have better intrinsic generalization also lead to a better recognition performance.**

I. Introduction

In recent years, principal component analysis (PCA) [12] has become a widely used tool for dimension reduction. One of the major limitations of PCA representation is that it is not able to capture statistics higher than the second order. Independent component analysis (ICA) [6] [1] (see [14] for a review) has been used to overcome this limitation by imposing statistical independence among the linear bases. By maximizing a discrimination measure among different classes, Fisher discriminant analysis (FDA) offers another popular linear subspace representation [8]. In computer vision, these representations have been applied to face recognition [27], [30], [3].

As the recognition performance of a classifier depends heavily on the choice of training data, it becomes important to study the generalization of a low dimensional representation through the equivalence relation it imposes on the image space. The importance of studying the equivalence classes for generalization is greatly emphasized by Vapnik [31]. In fact, Vapnik named the cardinality of equivalence classes a new concept (to be studied) for learning from small samples (p. 296, [31]). While the cardinality of equivalence classes is important for reducing dimensionality, for recognition performance, it is aslo important to study the properties of images in a particular equivalence class. Ideally, only images with similar underlying models should be grouped into an equivalence class. We will name this semantics-related aspect of generalization as *intrinsic generalization.* This isolates an intrinsic aspect of a representation that affects the recognition performance. Our study of intrinsic generalization for linear subspaces reveals that these representations group images from different models within the same equivalent class and are inherently sensitive to noise and deformations. By analyzing two problems with linear subspace representations, we propose a way to improve the intrinsic generalization of linear subspaces, the advantage of which is demonstrated by recognition result. We emphasize that our study is to characterize an important aspect of a representation through the study of intrinsic generalization, which leads to an important measure to compare different representations.

This paper is organized as follows. Section II gives a definition of generalization and then introduces intrinsic generalization and shows that of linear subspaces. Section III briefly describes the spectral representation of images, and spectral subspace analysis (SSA), and shows the intrinsic generalization of SSA through object synthesis. Section IV shows the experimental results on recognition of large datasets. Section V concludes the paper with a discussion on a number of related issues.

II. Intrinsic Generalization

In this paper, an image $\mathbf{I}$ is defined on a finite lattice $\mathcal{L} \subset \mathbf{Z}^2$, the intensity at pixel location $\vec{v} \in \mathcal{L}$ is denoted by $\mathbf{I}(\vec{v}) \in \mathcal{G} = [r_1, r_2]$, where r_1, r_2 bound the dynamic range of the imaging sensor, and Ω the set of all images on $\mathcal{L}$. A representation is a mapping defined as $f : \Omega \rightarrow R^K$. For a low dimensional representation, we require $K \ll |\mathcal{L}|$. Before we introduce the intrinsis generalization, here we first give a formal definition of generalization of representations.

A. Generalization of Low Dimensional Representations

In learning-based recognition of objects from images, a classifier/recognizer is often trained using some training data and is applied to classify/recognize future inputs in the form of test data. A key issue is to extend good performance on test inputs using information limited to the training set, commonly known as the generalization problem [4].

There are several ways of formulating the generalization problems and we have chosen the framework laid out by Bishop [4]. Let the observed images be generated from an unknown probability density P on Ω and the underlying true recognition function be denoted by $h : \Omega \mapsto \mathcal{A}$, where $\mathcal{A}$

[1] As pointed out in [26], ICA is a misnomer as estimated independent components are not guaranteed to be independent.

0-7803-7898-9/03/$17.00 ©2003 IEEE

is the set of all classes. For any classifier function g, its average generalization ability is defined as the probability that $g(\mathbf{I}) = h(\mathbf{I})$, i.e.

$$G(g) = Pr\{\mathbf{I}|\mathbf{I} \in \Omega, g(\mathbf{I}) = h(\mathbf{I})\}. \quad (1)$$

According to (1), obtaining good generalization becomes an optimization over all the possible classifier functions. In practice, since the underlying model $P(\mathbf{I})$ and $h(\mathbf{I})$ are generally unknown, several ways have been proposed. One way is to estimate $G(g)$ directly through a set separate from the training one with known class labels such as cross-validation [4]. Another way is to impose additional constraints on $G(g)$ based on some generic heuristics such as Akaike information criterion [1] and minimum description length [23], where a model with more free parameters is penalized. Yet another approach is to optimize an analytical bound based on statistical analysis such as the worst case performance of all the implementable classifiers of a neural network architecture [2]. Note that all the existing efforts on generalization have been focused on the generalization of classifiers.

Because of the high dimensionality of images, dimension reduction becomes necessary for computational reasons. In case of using a low dimensional representation f, the average generalization ability then becomes the probability that $\hat{g}(f(\mathbf{I})) = h(\mathbf{I})$ for an input $\mathbf{I}$ randomly sampled according to $P(\mathbf{I})$, where $\hat{g}$ is a classifier based on a low dimensional representation f. In other words, we have

$$\begin{aligned} G(\hat{g}, f) &= Pr\{\mathbf{I}|\mathbf{I} \in \Omega, g(f(\mathbf{I})) = h(\mathbf{I})\} \\ &= \textstyle\sum_{f(\mathbf{I})} Pr\{\mathbf{J}|\mathbf{J} \in \Omega, f(\mathbf{J}) = f(\mathbf{I})\} \\ &\quad 1_{\hat{g}(f(\mathbf{I}))=h(\mathbf{I})}, \end{aligned} \quad (2)$$

where $1_{x=y}$ is an indicator function, and we use $f(\mathbf{I})$ as the range of f on Ω. From (2), it is clear that f has a significant effect on the generalization of $\hat{g}$. Ideally, we want to group all the images from each class as a single equivalence class. (In this case, the classifier is trival.) While this is generally not possible for real applications, we want to group images from each class into a small number of equivalence classes, with each class having a large cardinality, as emphasized by Vapnik [31]. However, when making each equivalence class as large as possible, we do not want to include images from other classes, as this will make a good classification performance impossible. This leads to the need of analyzing equivalence class structures of low dimensional representations to achieve a good generalization performance.

B. Intrinsic Generalization

The previous analysis shows that the equivalence class structures of low dimensional representations are essential for a good generalization performance. In this paper, we focus on studying the images of a particular equivalence class through statistical sampling.

Definition: Given a representation f, the intrinsic generalization of an image $\mathbf{I}$ under f is defined as

$$S_I(\mathbf{I}) = \{\mathbf{J}|\mathbf{J} \in \Omega, f(\mathbf{J}) = f(\mathbf{I})\} \subset \Omega. \quad (3)$$

In other words, intrinsic generalization of image $\mathbf{I}$ includes all the images that cannot be distinguished from $\mathbf{I}$ under representation f. The recognition performance based on f depends critically on intrinsic generalizations of training images as the images in intrinsic generalizations are implicitly included in the training set. We define $S_I^0(\mathbf{I})$ as the images sharing the same underlying probability models with $\mathbf{I}$. Ideally, $S_I(\mathbf{I})$ should be as close as possible to $S_I^0(\mathbf{I})$. As $S_I^0(\mathbf{I})$ is generally not available, to explore $S_I(\mathbf{I})$, we employ statistical sampling through the following probability model:

$$q(\mathbf{J}, T) = \frac{1}{Z(T)} \exp\{-D(f(\mathbf{J}), f(\mathbf{I}))/T\}. \quad (4)$$

Here T is a temperature parameter, $D(.,.)$ a Euclidean or other distance measure, and $Z(T)$ is a normalizing function, given as $Z(T) = \sum_{\mathbf{J}\in\Omega} \exp\{-D(f(\mathbf{J}), f(\mathbf{I}))/T\}$. This model has been used for texture synthesis [33] and we generalize it to any representation. It is easy to see that as $T \to 0$, $q(\mathbf{J}, T)$ defines a uniform distribution on $S_I(\mathbf{I})$ [33]. The advantage of using a sampler is to be able to generate typical images in $S_I(\mathbf{I})$ so that $S_I(\mathbf{I})$ under f can be examined in a statistical sense.

C. Intrinsic Generalization of Linear Subspaces

Linear subspace representations of images, including PCA, ICA, and FDA, assume that f is a linear map, and $S_I(\mathbf{I})$ forms a linear subspace. While these methods are successful when applied to images belonging to a specific nature, e.g. face images, their generalization seems poor if we consider $S_I(\mathbf{I})$ under these linear subspace methods in Ω.

If $S_I^0(\mathbf{I})$ is available then one can analyze the overlap between the sets $S_I^0(\mathbf{I})$ and $S_I(\mathbf{I})$. If not, then one has to resort to some indirect technique such a random sampling to compare elements of the two sets. Random sampling seems sufficient in that the typical images $S_I(\mathbf{I})$ are very different from $\mathbf{I}$. To illustrate these ideas, we have used a PCA of the ORL face dataset[2], which consists of 40 subjects with 10 images each; we have obtained similar results using other linear subspaces. We calculate the eigen faces corresponding to the 50 largest eigenvalues. Under PCA, given an image $\mathbf{I}$, $f(\mathbf{I})$ is the projections of $\mathbf{I}$ along eigen faces. We define the reconstructed image of $\mathbf{I}$ as $\pi(\mathbf{I}) = \sum_{i=1}^{K} < \mathbf{I}, \mathbf{V}_i > \mathbf{V}_i$, where $\mathbf{V}_i$ is the ith eigen face and $< .,. >$ is the canonical inner product. Fig. 1(a) shows a face image in the dataset and Fig. 1(b) shows the reconstructed image with $K = 50$. We then use a Gibbs sampler to generate samples of $S_I(\mathbf{I})$. Fig. 1(c)-(f) show four samples of $S_I(\mathbf{I})$. (For Fig. 1(f), the object in the middle is used as boundary condition, i.e., pixels on the object are not updated) In other words, these images have the same 50 eigen decomposition. Note that $S_I(\mathbf{I})$ is defined on Ω and these images are far from each other in Ω. As expected, the corresponding reconstructed images are identical to Fig. 1(b).

Because $S_I(\mathbf{I})$ consists of images from various underlying probability models, the linear subspace representations can

[2] http://www.uk.research.att.com/facedatabase.html.

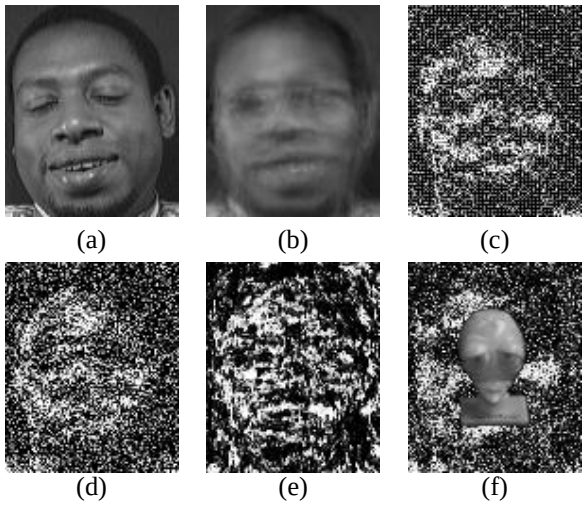

Fig. 1. (a) A face image. (b) Reconstructed image using $K = 50$ principal components. (c)-(f) Four random samples from the set $S_I(\mathbf{I})$, with $\pi(\mathbf{I})$ identical to the one shown in (b).

make the subsequent classification intrinsically sensitive to noise and other deformations. To show that, Fig. 2(a) gives three different face images which share the exactly same eigen representation (bottom row). On the other hand, Fig. 2(b) shows three similar images whose eigen representations correspond to three different faces.

We emphasize here that the sampling is very from reconstruction. The sampling is to draw a typical sample from the set of all the images with a particular low dimensional representation while reconstruction gives one in the set whose coefficients are zero along the dimensions complement to the given subspace. To illustrate this, Fig. 3 shows an example of a one-dimensional subspace in a two-dimensional space. In this case, the reconstructed "image" of "x" is the point given by "+" in Fig. 3(a) while the sampling can return any point with equal probability along the solid line shown in Fig. 3(b). This shows clearly that the reconstructed image may not provide much information about all the other images having the same low dimensional representation.

These results, while generated based PCA, are valid to an arbitrary linear subspace since the sampling tries to match the representation. The main problem of linear subspace representations, as revealed here, is that these representations can not take into account that most images in the image space are white noise images.

III. Spectral Subspace Analysis

A. Spectral Representation of Images

As discussed earlier, an ideal representation f for $\mathbf{I}$ will be such that $S_I(\mathbf{I}) = S_I^0(\mathbf{I})$. There are two important limitations of the linear methods that need to be addressed: (i) As the vast majority images in Ω are white noise images, a good approximation of $S_I^0(\mathbf{I})$ for an image of object(s) must handle white noise images effectively; otherwise, $S_I(\mathbf{I})$ will concentrate on white noise images. Experiments shows that linear representations suffer from this problem. (ii) Another issue is the linear superposition assumption, where each basis

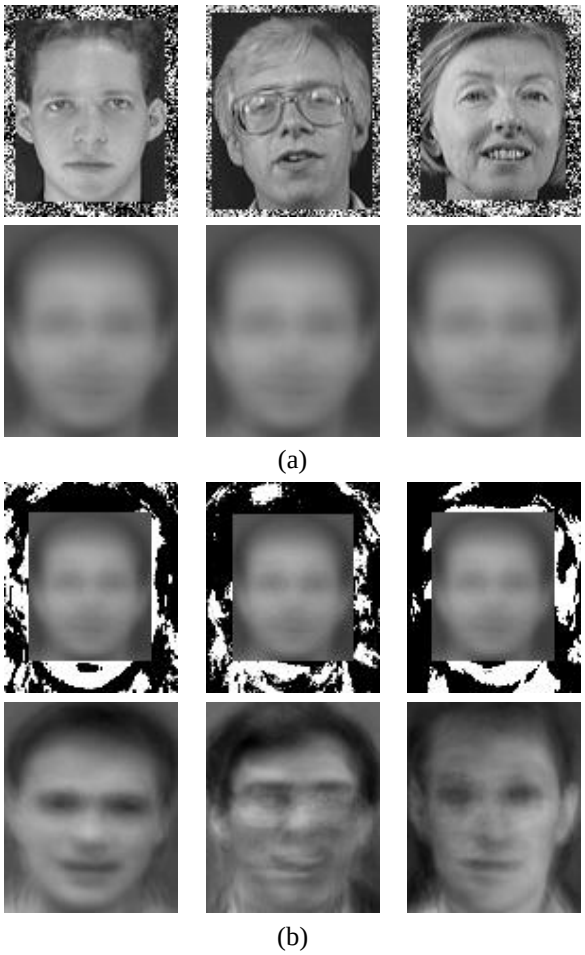

Fig. 2. Examples of different images with identical eigen decompositions and similar images with different eigen decompositions. The top row shows the images and the bottom reconstructed. (a) Three different images with the same eigen representations. (b) Three similar images with different eigen representations.

contributes independently to the image. In contrast, pixels on objects are dependent and efficient models should exploit this dependency.

The issue of white noise images can be dealt with effectively through the method of types [7] as the white noise images are grouped together under types. However, the direct use of types does not provide enough constraints as only the histogram of images is used. We generalize the type definition by including marginals of filter responses (of the input image) with respect to a set of filters, which also incorporates local pixel dependence through filtering.

The representation of using marginals of filtered images can be justified in many ways: (i) by assuming that *small disjoint regions in the frequency domain are independent*. That is, partition the frequency domain into small disjoint regions and model each region by its marginal distribution. The partitioning of the frequency also leads to spatial filters. (ii) Wavelet decompositions of images are local in both space and frequency, and hence, provide attractive representations for objects in the images. We convolve an image with the filters and compute the marginals. Each image is then represented by

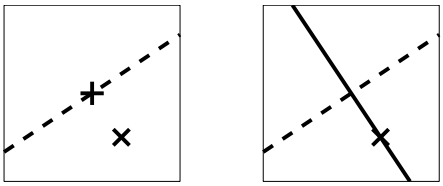

Fig. 3. An illustration example of the difference between sampling and reconstruction. Here the dashed line represents a one-dimensional subspace in a two-dimensional space. For a training example (marked as 'x'), the sampling is to draw a random point along the solid line in (b) while the reconstructed image is a single point given by '+' in (a).

a vector consisting of all the marginal distributions. We shall call this representation *spectral representation*, each of these vectors a *spectral histogram*, and the set of all valid vectors the *spectral space*. Elements of a spectral histogram relate to the image pixels in a nonlinear fashion, and hence, avoid the linearity issue mentioned earlier.

This representation has also been suggested through psychophysical studies on texture modeling [5], and has been used in the texture modeling and synthesis [11], [34], [18] and texture classification [19]. Both the histogram of input images [29] and joint histograms of local fields [25] have been used for object recognition.

B. Spectral Subspace Analysis

In this method, the strategy is to first represent each image in the spectral space and then apply a linear subspace method, such as PCA, ICA or FDA, in the spectral histogram space[3]. Name these corresponding methods as SPCA, SICA, and SFDA, and call them collectively as spectral subspace analysis (SSA).

To demonstrate the effectiveness of SSA representations, we explore their intrinsic generalization through sampling. As in the linear subspace case, we use SPCA for experiments; similar results have been obtained using other linear spectral subspaces.

First, bases in the spectral space are computed based on training images. Given an image, its spectral representation is computed and then projected onto a spectral subspace. We use a Gibbs sampling procedure to generate images that share the same spectral representation. Fig. 4 shows two sets of examples; Fig. 4(a) shows three texture images and Fig. 4(b) shows three objects. These examples show that the spectral subspace representation captures photometric features as well as topological structures, which are important to characterize and recognize images.

IV. Experimental Results for Recognition

To demonstrate the effectiveness of SSA representations, we use several data sets and compare their performance with that of linear subspace representations. In our experiments, the number of principal components is determined by thresholding the ratio of a component's eigenvalue and the largest eigenvalue. (If the same threshold is applied to PCA, it tends to keep more components as the dimension of the input space here is much larger). We have used the same number of components for ICA, FDA, SICA, and SFDA as PCA and SPCA. Here ICA is calculated using the FastICA algorithm [13] and FDA based on an algorithm by Belhumeur et al. [3]. We use the nearest neighbor classifier for recognition. To calculate the spectral histogram, we use a fixed set of 21 filters. These filters were chosen automatically from a larger set using a filter selection algorithm [17] for the ORL face dataset.

[3]Note a reconstructed spectral histogram may be outside the spectral space and here we ignore this complication.

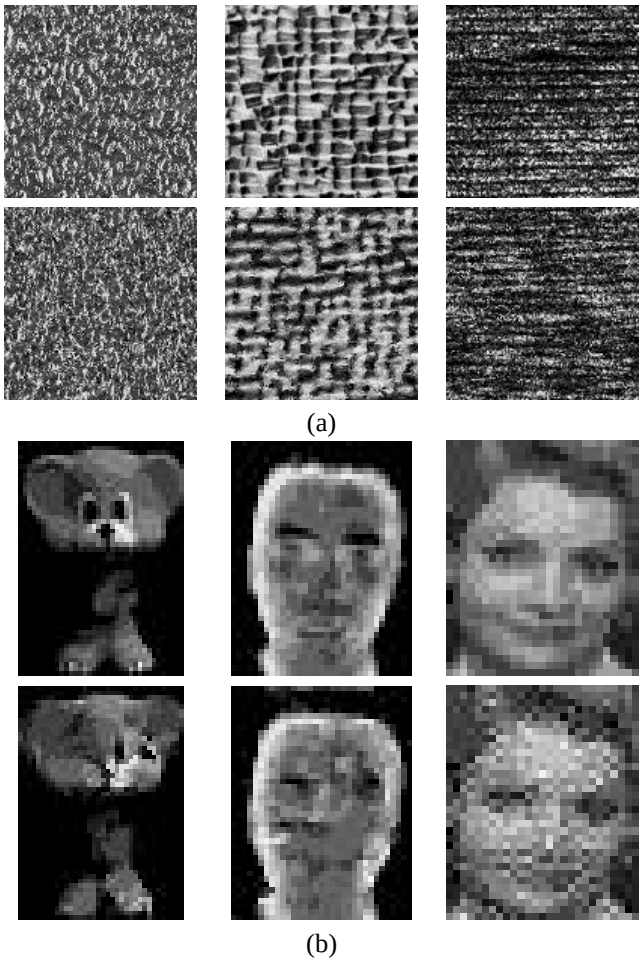

Fig. 4. Samples from SPCA intrinsic generalization. In each panel the top row shows the input image and the bottom a typical sample from its intrinsic generalization. (a) Three textures. (b) One object and one face image. Boundary conditions need to be taken with care when sampling from $S_I(\mathbf{I})$.

A classifier's performance in a low dimensional space depends on intrinsic generalization of the representation to the test data. The result for a new image of a classifier is determined by the decision region partitions in the feature space and thus in the image space. Given a training set B, we define the extrinsic generalization set of $\mathbf{I} \in B$ as:

$$\begin{aligned} S_E(\mathbf{I}) \quad = \quad & \{\mathbf{J} | \mathbf{J} \in \Omega, L(\mathbf{J}|B) = L(\mathbf{I})\} \ \setminus \\ & \{\mathbf{J} | \mathbf{J} \in \Omega, L(\mathbf{J}|B \setminus \mathbf{I}) = L(\mathbf{I})\}. \end{aligned} \tag{5}$$

Here $L(\mathbf{I})$ is the label of $\mathbf{I}$ and $L(\mathbf{J}|B)$ is the label of image

$\mathbf{J}$ assigned by the classifier given training set B. To separate the effectiveness of a representation from that of the choice of training and test data, we have also used (uniformly) randomly generated bases, which we call random component analysis (RCA) and spectral random component analysis (SRCA).

First we use the Columbia Object Image Library (COIL-100)[4] dataset, which consists of images of 100 3-D objects with varying pose, texture, shape and size. Pontil and Verri [22] applied SVM (Support Vector Machines) method to 3D object recognition. Yang et al. [32] proposed a learning algorithm, named SNoW (Sparse Network of Winnows) for appearance based recognition and applied their learning algorithm to the full COIL dataset and compared with SVM methods. They tested their method by varying the number of training views. For a fair comparison with the results in [32], these color images were converted to gray level images and downsampled to size 32×32, which have been used in all the experiments here.

As in [32], we vary the number of training views per object. Tab. I shows the recognition rates on the dataset using PCA, ICA, FDA, SNoW [32] and Linear SVM [32] as well as SSA methods. While the COIL-100 dataset is considered to be easy with enough training data, Tab. I reveals clearly the generalization of different representations. Under all the conditions, SSA methods outperform the other methods. Among the SSA methods, SFDA gives the best performance. However, FDA in the image space does not outperform other linear methods; this is because these images are not linearly separable in the image space and linear FDA bases are not effective. This result is consistent with that of [20], which showed that FDA may not outperform PCA in general. Another interesting point of Tab. I is that RCA and SRCA give comparable results to those of other bases, suggesting that the choice of commonly used different bases within a space may not be that critical for recognition as none of them is considerably better than a random one in term of recognition performance.

TABLE I
RECOGNITION RESULTS FOR THE COIL-100 DATASET

Methods	Training / test per object			
	36 / 36	18 / 54	8 / 64	4 / 68
PCA	98.58%	96.67%	87.23%	75.82%
ICA	98.47%	96.52%	87.91%	76.03%
RCA	98.61%	96.30%	86.95%	75.35%
FDA	97.61%	92.63%	82.13%	56.82%
SNoW [32]	95.81%	92.31%	85.13%	81.46%
Linear SVM [32]	96.03%	91.30%	84.80%	78.50%
SPCA	99.39%	97.13%	89.25%	82.91%
SICA	99.39%	97.03%	89.23%	82.72%
SRCA	99.42%	96.89%	89.13%	82.95%
SFDA	99.91%	98.8%	94.44%	87.37%

We have also applied SSA methods to face and texture datasets. To demonstrate more convincingly that the SSA representations are able to represent different types of images at the same time, we create a dataset by combining ORL face dataset, a texture dataset, and the COIL-100 dataset. The resulting dataset consists of 180 different classes with 40 textures, 100 objects, and 40 faces and a total of 10160 images. To measure the reliability of the recognition result, for each test image $\mathbf{I}$, we calculate $p_0(i|I)$ as

$$p_0(i|\mathbf{I}) = \frac{\exp\{\frac{-D(f(C^{(i)}),f(\mathbf{I}))}{\min_j D(f(C^{(j)}),f(\mathbf{I}))}\}}{\sum_k \exp\{\frac{-D(f(C^{(k)}),f(\mathbf{I}))}{\min_j D(f(C^{(j)}),f(\mathbf{I}))}\}}, \quad (6)$$

where $C^{(i)}$ is the training set of category i, $D(f(C^{(i)}), f(\mathbf{I}))$ the minimum distance between the representation of $\mathbf{I}$ and all the images in $C^{(i)}$. We calculate the entropy of $p_0(i|\mathbf{I})$ as a measure of reliability. Tab. II shows the recognition result along with the average entropy of $p_0(i|\mathbf{I})$ over all test images. Again, all the SSA methods outperform linear subspace methods. In addition, the average entropy of the SSA methods is much lower than that of the linear subspace methods.

TABLE II
RECOGNITION RATE AND THE AVERAGE ENTROPY OF $p_0(i|\mathbf{I})$ FOR THE COMBINED DATASET

Methods	Total training / test images			
	5080 / 5080		1300 / 8860	
	Recog. rate	Average entropy	Recog. rate	Average entropy
PCA	77.91%	1.69 bits	71.60%	2.61 bits
ICA	77.83%	1.72 bits	72.31%	2.55 bits
RCA	76.52%	1.99 bits	69.82%	3.08 bits
FDA	76.36%	2.72 bits	72.83%	2.83 bits
SPCA	98.31%	0.62 bit	91.73%	1.36 bits
SICA	98.41%	0.55 bit	91.22%	1.28 bits
SRCA	97.54%	0.79 bit	90.59%	1.54 bits
SFDA	99.65%	0.60 bit	96.55%	1.63 bits

V. DISCUSSION

One of the major obstacles of developing a generic vision system is the generalization of an underlying representation. By studying the intrinsic generalization of a representation, we can better understand and predict its performance under different conditions. To our knowledge, this is the first attempt to provide a quantitative generalization measure intrinsic to a representation; in contrast, generalization is commonly tied to recognition performance, which depends on the choice of the classifier and the choice of training and test data. Our study on the intrinsic generalization of linear subspace representations in the image space shows that they cannot generalize well as images from different models tend to be grouped into one equivalence class; we emphasize that this result holds for any low dimensional linear subspace in the image space. We have suggested a way to improve the intrinsic generalization by implementing linear subspaces in the spectral space. We have demonstrated substantial improvement in recognition on large datasets.

However, our goal is not to show that SSA representation is optimal in general. In fact, if classes consist of white noise

[4]Available at http://www.cs.columbia.edu/CAVE.

like images, SSA representations would be very ineffective. Rather our emphasis is on the importance of the underlying representation for object images. An ideal representation of image $\mathbf{I}$ is $S_I^0(\mathbf{I})$, which can be implemented only when the true underlying object models and the physical imaging process are available; this leads to the analysis-by-synthesis paradigm [9]. When $S_I^0(\mathbf{I})$ is not available explicitly, one needs to approximate it. There is a trivial solution for a good approximation by forcing $S_I(\mathbf{I}) = \{\mathbf{I}\}$. However, the generalization is very poor and it requires literally all possible images in the training set. A good representation should approximate $S_I^0(\mathbf{I})$ well and $|S_I(\mathbf{I})|$ should be as large as possible. These two constraints provide the axes of forming a continuous spectrum of different representations and allow us to study and compare them. For example, only marginal distributions are used in the spectral representation; one can describe and synthesize $\mathbf{I}$ better by incorporating joint statistics [24]; however, this obviously decreases $|S_I(\mathbf{I})|$. Within linear subspace methods, one can also decrease $|S_I(\mathbf{I})|$ by imposing additional constraints on bases and coefficients, such as the non-negative constraints [16][5]. Due to the complexity of $S_I^0(\mathbf{I})$, a very close approximation using some low dimensional representations may not be feasible. An alternative is to combine the analysis-by-synthesis paradigm [9] and a low dimensional representation based approach. The hypothesis pruning by Srivastava et al. [28] provides such an example, where a low dimensional representation selects plausible hypotheses for a analysis-by-synthesis model. In this framework, the difference among low dimensional representations is their effectiveness of selecting good hypotheses rather than providing a final answer.

Acknowledgments This research was supported in part by grants NIMA NMA201-01-2010, NSF DMS-0101429, and ARO DAAD19-99-1-0267. D.L.W. was supported in part by an NSF grant (IIS-0081058) and an AFOSR grant (F49620-01-1-0027). The authors would like to thank the producers of the COIL, ORL, and texture datasets for making them publicly available.

REFERENCES

[1] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in B. N. Petrov and F. Csaki (eds.), *2nd International Symposium on Information Theory*, pp. 267–281, 1973.

[2] E. B. Baum and D. Haussler, "What size net gives valid generalization?" *Neural Computation*, vol. 1, no. 1, pp. 151–160, 1989.

[3] P. N. Belhumeur, J. P. Hepanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19(7), pp. 711–720, 1997.

[4] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, UK, 1995.

[5] C. Chubb, J. Econopouly, and M. S. Landy, "Histogram contrast analysis and the visual segregation of IID textures," *J. Opt. Soc. Am. A*, vol. 11, pp. 2350–2374, 1994.

[6] P. Comon, "Independent component analysis, A new concept?" *Signal Processing*, vol. 36(4), pp. 287-314, 1994.

[7] I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York, 1981.

[5]Rigorously speaking, the bases with non-negative constraints do not form linear subspaces anymore.

[8] R. A. Fisher, "The use of multiple measures in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179-188, 1936.

[9] U. Grenander, *General Pattern Theory*, Clarendon Press, Oxford, 1993.

[10] U. Grenander and A. Srivastava, "Probability models for clutter in natural images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23(4), pp. 424-429, 2001.

[11] D. J. Heeger and J. R. Bergen, "Pyramid-based texture analysis/synthesis," in *Proceedings of SIGGRAPHS*, pp. 229–238, 1995.

[12] H. Hotelling, "Analysis of a complex of statistical variables in principal components," *Journal of Educational Psychology*, vol. 24, pp. 417-441, 498-520, 1933.

[13] A. Hyvärinen, "Fast and robust fixed-point algorithm for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10(3), pp. 626–634, 1999.

[14] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, Wiley-Interscience, 2001.

[15] M. Lades, J. C. Vorbruggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Wurtz, and W. Konen, "Distortion invariant object recognition in the dynamic link architecture," *IEEE Transactions on Computers*, vol. 42, pp. 300-311, 1993.

[16] D. D. Lee and S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[17] X. W. Liu and D. L. Wang, "Appearance-based recognition using perceptual components," in *Proceedings of the International Joint Conference on Neural Networks*, vol. 3, pp. 1943–1948, 2001.

[18] X. Liu and D. L. Wang, "A spectral histogram model for texton modeling and texture discrimination," *Vision Research*, vol. 42, no. 23, pp. 2617–2634, 2002.

[19] X. Liu and D. L. Wang, "Texture classification using spectral histograms," *IEEE Transactions on Image Processing*, in press, 2003.

[20] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23(2), pp. 228-233, 2001.

[21] S. K. Murase and S. K. Nayar, "Visual learning and recognition of 3-d objects from appearance," *International Journal of Computer Vision*, vol. 14, pp. 5-24, 1995.

[22] M. Pontil and A. Verri, "Support vector machines for 3D object recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20(6), pp. 637-646, 1998.

[23] J. Rissanen, "Modelling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[24] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelets," *International Journal of Computer Vision*, vol. 40(1), pp. 49–71, 2000.

[25] B. Schiele and J. L. Crowley, "Recognition without correspondence using multidimensional receptive field histograms," *International Journal of Computer Vision*, vol. 36, pp. 31–50, 2000.

[26] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annual Review of Neuroscience*, vol. 24, pp. 1193-1216, 2001.

[27] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *J. Opt. Soc. Am. A*, vol. 4(3), pp. 519-524, 1987.

[28] A. Srivastava, X. W. Liu, and U. Grenander, "Universal analytical forms for modeling image probabilities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23(9), 2002.

[29] M. J. Swain and D. H. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, 11-32, 1991.

[30] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, pp. 71-86, 1991.

[31] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed., Springer-Verlag, New York, 2000.

[32] M. H. Yang, D. Roth, and N. Ahuja, "Learning to recognize 3D objects with SNoW," in *Proceedings of the Sixth European Conference on Computer Vision*, vol. 1, pp. 439-454, 2000.

[33] S. C. Zhu, X. W. Liu, and Y. Wu, "Exploring Texture ensembles by efficient Markov chain Monte Carlo," *IEEE Transactions on Pattern Recognition and Machine Intelligence*, vol. 22, pp. 554-569, 2000.

[34] S. C. Zhu, Y. N. Wu, and D. Mumford, "Minimax entropy principle and its application to texture modeling," *Neural Computation*, vol. 9, pp. 1627–1660, 1997.