Independent Spectral Representations of Images for Recognition

Xiuwen Liu and Lei Cheng

Department of Computer Science

Florida State University

Tallahassee, Florida 32306-4530, USA

{*liux, leicheng*}*@cs.fsu.edu*

http://fsvision.fsu.edu

Revised Version, Submitted to the special issue on *Bayesian and Statistical Approaches of Vision*, *Journal of the Optical Society of America, A*.

In recent years, studies have shown that independent components of local windows of natural images resemble the receptive fields of cells in the early stages of the mammalian visual pathway. However, the role of the independence in visual recognition is not well understood. In this paper, we argue that the independence resolves the curse of dimensionality by reducing the complexity of probability models to the linear order of the dimension. In addition, we show empirically that the complexity reduction does not deteriorate the recognition performance on all the datasets we have used based on an independent spectral representation. In this representation, an input image is first decomposed into independent channels given by the estimated independent components from training images and each channel's response is then summarized using its histogram as an estimate of the underlying probability model along that dimension. We demonstrate the sufficiency of the proposed representation for image characterization by synthesizing textures and objects through sampling and for recognition by applying it to large datasets. Our comparisons show that the independent spectral representation often gives better recognition performance.

OCIS codes: 100.5010,100.2960,330.4060,330.6110

1. Introduction

Understanding vision has evolved to be one of the most fundamental problems in understanding intelligence and in designing intelligent machines.[1] Since Marr's influential paradigm of computational vision,[2] it has been widely accepted that vision can be understood as an information processing system, where Bayesian inference[3] has become a dominating approach. In the Bayesian framework, vision is formulated as a statistical inference problem:

$$P(S|I) = \frac{P(I|S)P(S)}{P(I)} \propto P(I|S)P(S), \tag{1}$$

where I is an input image or image sequence in general and S represents plausible interpretations. For an input image, the interpretations that a vision system should produce are the ones with high posterior probability $P(S|I)$, a principle dated back to Helmholtz.[4] While this formulation provides a formal framework to formulate vision problems in a principled way and many of the classical ones have been elegantly reformulated in this framework (see Knill and Richards[3] and references therein for examples), it does not, however, imply a framework to implement an efficient computational vision system. The curse of dimensionality[5] and the lack of sufficient probability models for vision problems have seriously limited the Bayesian framework's applicability to complex real world problems.

The curse of dimensionality[5] is a widely observed phenomenon in computational modeling; in the context of Bayesian framework it refers to that both representing and estimating high dimensional probability models require resources that grow exponentially with the dimension (see Section 2 for further explanations). As a result, this formal framework is often twisted to incorporate unprincipled rules and knowledge to expedite the inference process. While impressive results can be obtained, this twist seems to invalidate the formality of the framework itself. Given the curse of dimensionality, it seems that the only computationally feasible approach to vision is to create sufficient models that are not subject to the curse of dimensionality. A plausible solution is to reduce the dimensions through some dimension reduction techniques, such as principal component analysis[6–8] and impose probability models in the reduced dimensions. Given the great variability in natural images, while the number of dimensions can be reduced significantly within one class of images, such as faces, the number of dimensions required to model different classes of images seems still too high to be computationally feasible. Independence[9] offers an alternative and general principle to overcome the problem, the idea of which has been proposed in the context of projection pursuit[10] and unsupervised learning.[11] Under the independence assumption, the complexity of probability models does not grow exponentially and thus more dimensions can be handled computationally. Studies in this direction have been fruitful in relating independent components of natural images to the properties of the receptive fields of cells in the early stages of the mammalian visual pathway (see e.g. Field,[12,13] Olshausen and Field,[14] and Bell and Sejnowski;[15] Simoncelli and Olshausen,[16] and Srivastava et al.[17] for reviews). In general, independent components are estimated using local windows (of a specified size) sampled from a set of natural images and thus the independence is a property imposed on the ensemble of the training image windows of a specified size. Given that the primary role of a vision system is to perform recognition, i.e., to differentiate objects based on the images, a critical question is how to relate the independence to the recognition performance. Recently, the significance of these studies has been questioned based on the argument that the principles used in deriving those properties may not play an important role for visual recognition. Vasconcelos and Carneiro[18] demonstrated empirically using image retrieval experiments that independent components often give worse performance than commonly used bases.

In this paper, by viewing images as realizations sampled from the underlying probability distribution, we argue that the linear representation given by projecting the input onto the independent bases is not statistically meaningful. This perspective requires that an image be characterized by the underlying probability models, not directly by the pixel values in the input. This leads us to propose an independent spectral representation. In this representation, independent components estimated from input image windows are used as filters, which are called *independent filters*, to differentiate this view from that as a set of linear bases of the entire input images. This is also consistent with studies on comparing independent components with the receptive fields in the visual system.[15] In analogy to the spatial/frequency channels in the visual system,[19,20] an independent filter also gives rise to an independent channel, the response of which is modeled as the convolution of the input with the independent filter. The response from each independent channel is represented by its histogram as an estimation of the underlying probability model along this independent dimension. We systematically demonstrate the sufficiency of the proposed representation through synthesizing textures and objects. Then we show through recognition experiments on texture classification and face recognition that the independent spectral representation produces the best results on all the datasets we have used.

The paper is organized as follows. Section 2 introduces independent components of image windows and the independent spectral representation of images. Section 3 uses texture and object synthesis to demonstrate the sufficiency of the proposed representation. Section 4 shows our experimental results on texture, face, and infrared face datasets. Section 5 discusses a number of issues and Section 6 concludes the paper.

2. Independent Spectral Representations of Images

A. Independent Filters of Images

In image analysis, it is often desirable to represent the input image in a (sub)space to reveal some of the characteristics of the data and reduce the dimension, where linear representation is often used due to its computational and analytical properties. Under the linear representation, an observed image window $\mathbf{I}$ of size $n \times m = N$, represented as a vector $\vec{y} = (I_{1,1}, I_{1,2}, \ldots, I_{n,m})' = (y_1, \ldots, y_N)'$, is assumed to be generated by a linear combination of hidden but fixed K factors $\mathbf{S} = (\vec{S}_1, \ldots, \vec{S}_K)$. Here $'$ indicates transpose operation and each S_i is a vector of length N. This leads to the following linear generative model:

$$\vec{y} = \sum_{i=1}^{K} x_i \vec{S}_i = \mathbf{S}\vec{x}, \qquad (2)$$

where $\vec{x} = (x_1, \ldots, x_K)'$ is called the representation of $\vec{y}$ under $\mathbf{S}$. Under the linear assumption, recovering the representation given an input is through pseudo inverse, given by,

$$\vec{x} = \mathbf{W}\vec{y}. \tag{3}$$

We assume each column of $\mathbf{S}$ and each row of $\mathbf{W}$ have a unit length. In the probabilistic modeling of images, $\mathbf{S}$ and $\mathbf{W}$ are estimated from a set of training images $\vec{y}^{(1)}, \ldots, \vec{y}^{(M)}$. These training images are assumed to be samples from a model with the same underlying factors. More formally, each y_i is assumed to be a random variable and thus each x_i is also a random variable. Each training image $\vec{y}^{(m)}$ is viewed as one sample in the joint space given by $y_1, \ldots, y_N$. By imposing different statistical properties on the space given by $\vec{x}$, different techniques have been proposed to estimate $\mathbf{S}$ and $\mathbf{W}$ from the training set. For example, principal component analysis[6,7] captures the most variance in the space given by $\vec{x}$. The principal components are orthonormal and thus $\mathbf{W} = \mathbf{S}'$. One of the major limitations of principal components is that they can only capture second-order statistics. Higher order statistics, which are shown to be important for effective natural image representations,[14] are ignored in principal components. To overcome this limitation, independent component analysis[9] is proposed, where $\mathbf{S}$ and $\mathbf{W}$ are estimated by minimizing the statistical dependence of the resulting representation in the given set. Formally, we have,

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} KL(p(\vec{x}), p(x_1) \cdots p(x_K)), \tag{4}$$

where $\vec{x} = \mathbf{W}\vec{y}$ as in Eq. (3), KL is the Kullback-Leibler divergence, $p(\vec{x})$ the joint distribution, and $p(x_i)$ the ith marginal distribution. The right term in Eq. (4) is the mutual information among $x_1, \ldots, x_K$. Here $p(\vec{x})$ and $p(x_i)$ are estimated from training images $\vec{y}^{(1)}, \ldots, \vec{y}^{(M)}$. The direct calculation of independent components for a large K is computationally not feasible as it requires to estimate the full joint distribution and evaluate the mutual information. In practice, many algorithms have been proposed by optimizing some criteria related to statistical independence (see

Hyvärinen[21] for a survey).

In this paper, the independent components of a specified window size are computed using the FastICA algorithm by Hyvärinen[22] and the rows of the estimated $\mathbf{W}$ are used as independent filters. The algorithm was derived based on a first order approximation of negentropy, which can be used to evaluate mutual information.[9] This iterative algorithm was empirically shown to be 10 to 100 times faster than many ICA algorithms.[22]

To calculate the independent filters, first 40,000 image windows of a specified size from 3×3 to 25×25 are extracted from the training images. These image windows are used as observations and the FastICA algorithm is applied on the set to estimate $\mathbf{S}$ and $\mathbf{W}$ at the same time. Fig. 1 shows independent components estimated using samples from a texture dataset with different window sizes. Note that $\mathbf{S}$ and $\mathbf{W}$ in independent component analysis are not in general orthonormal; to be computationally efficient, we have also orthonormalized $\mathbf{W}$ by $\mathbf{W}_o = (\mathbf{W}'\mathbf{W})^{-1/2}\mathbf{W}$. These filters resemble orientation-sensitive derivative filters. Note that the independent filters depend on the window size and do not match analytical filters, even though some of the independent filters may have a similar shape with some analytical filters, such as Gabor filters.

Assuming that the resulting representations of the given filters are statistically independent, the Kullback-Leibler distance between two joint distributions is the sum of their corresponding marginal distributions as shown by the following equation:

$$\begin{aligned}
&KL(p_1(x_1, \cdots, x_K), p_2(x_1, \cdots, x_K)) \\
&= \int_{x_1} \cdots \int_{x_K} p_1(x_1, \cdots, x_K) \log \frac{p_1(x_1, \cdots, x_K)}{p_2(x_1, \cdots, x_K)} dx_1 \cdots dx_K \\
&= \int_{x_1} \cdots \int_{x_K} p_1(x_1) \cdots p_1(x_K) \sum_{i=1}^{K} \log \frac{p_1(x_i)}{p_2(x_i)} dx_1 \cdots dx_K \\
&= \sum_{i=1}^{K} \int_{x_i} p_1(x_i) \log \frac{p_1(x_i)}{p_2(x_i)} dx_i \\
&= \sum_{i=1}^{K} KL(p_1(x_i), p_2(x_i))
\end{aligned} \tag{5}$$

where $p_i(x_1, \cdots, x_K)$ is the joint distribution and $p_i(x_j)$ the jth marginal distribution. Equation

(5) implies that under the independence assumption we can compare probability models of images by comparing the marginal distributions.

Compared to the joint representation, the marginal distribution representation provides a low complexity representation without loss of information if the independent filters are completely independent. Equation (5) shows a distinctive advantage of an independent representation. That is, it greatly reduces the structural complexity of representing the probability distributions. To illustrate, suppose that probability distributions are represented discretely using L_i bins along dimension x_i. To represent the joint distribution of K dimensions, we need $\prod_{i=1}^{K} L_i$ cells. For example, if $K = 11$ and $L_i = 10, \forall i$, it would require 10^{11} cells to represent the joint distribution, which would use roughly all the neurons in the brain! On the other hand, representing all the marginals requires only $\sum_{i=1}^{K} L_i$ cells. For the above example, we only need 110 cells. In addition, the number of samples needed to estimate the distributions is also reduced significantly. To emphasize this significant advantage of independent representations, Fig. 2 shows the number of cells needed for a joint distribution and the corresponding marginal distributions. It is clear that independent components resolve the curse of dimensionality, which is a fundamental factor in determining whether a successful method for low dimensions can be extended to high dimensional cases.

B. Independent Spectral Representations

It is clear from the definition that independent filters and the corresponding representations are derived based on a set of unlabeled images and the independence property is valid on the ensemble of the images used in estimating independent components. Because of this setting, a direct use of independent components on a single image window is not statistically meaningful. Given that the primary goal of a vision system is to differentiate desirable objects such as food sources from dangerous objects, distinguishing and recognizing different natural images is clearly critical. A plausible visual representation therefore must be reasonably accurate for recognition. This raises a fundamental question for independent components. To be useful, a representation based on independent components must be derived, which can be used to effectively distinguish among objects associated with natural images.

Given a large image of an object, by breaking the image into overlapping windows of a specified size, we estimate the marginal distributions using the representations of these windows and thus the underlying joint distribution. If we ignore the spatial dependence of the image windows, we can use the empirical histogram of the representations of the image windows as the maximum likelihood estimate of the corresponding marginal distribution.[23] Also the convolution with a particular image window is the dot product between the window and the mirrored filter and thus Eq. (3) can be implemented using convolution. These observations lead to the following representation of an image. Given a set of independent filters, we convolve the input image with the filters. The histogram of the filtered image is computed which gives an estimate of the marginal distribution along the dimension given by the filter. As shown in Eq. (5), the marginal distributions along different dimensions together represent the underlying joint probability distribution under the independence assumption. Thus the histograms of filtered images provide a characterization of the input image. In the context of spectral representation,[24] we call this representation *independent spectral representation* as the filters are assumed to be independent. Note that the independent spectral representation is conceptually different from commonly used linear representations where Eq. (3) is viewed as a means for dimension reduction (e.g. Sirovich and Kirby[8] and Socolinsky and Selinger[25]). In the latter one, the recognition is done based on the projection of each image as a whole. While it can be effective for recognition on some datasets (see Section 4 for examples), it is not consistent with the probabilistic view of image modeling.

The use of histograms of filtered images has been proposed for texture modeling. Motivated by psychophysical studies,[26,27] Marginal distributions of filtered images have been used effectively for texture synthesis[28–30] and texton and texture discrimination.[24] They have also been extended empirically to object recognition [31,32] and segmentation.[33] While these representations rely the independence assumption implicitly, the assumption was not formulated formally.

3. Texture and Object Synthesis Using Independent Spectral Representations

From its definition, the independent spectral representation represents the object in an image using probability models of local windows, which ignores the positional sensitive information in the image, resulting in translation invariant feature statistics. While it is desirable for cases where global topological structures are not important such as textures, it seems that the independent spectral representation may be an oversimplified representation that is only effective for certain types of images.

To address this issue and demonstrate the sufficiency of the proposed representation, we utilize texture and object synthesis to systematically generate images that share a given independent spectral representation. Given a spectral representation $(p_{I^{obs}}(x_1), \ldots, p_{I^{obs}}(x_K))$ estimated from an observed image I^{obs}, the synthesis is to draw elements randomly from the set:

$$\Omega(I^{obs}) = \{I | p_I(x_i) = p_{I^{obs}}(x_i), i = 1, \ldots, K\}.$$

Here we use $p_I(x_i)$ to indicate that $p(x_i)$ is estimated from image I. An exhaustive search among all the possible images is computationally not feasible. We utilize a sampling procedure by inducing the following probability model on all images,

$$q(I|T, p_{I^{obs}}(x_1), p_{I^{obs}}(x_K)) \propto \exp(-\sum_{i=1}^{K} D(p_I(x_i), p_{I^{obs}}(x_i))/T), \tag{6}$$

where T is a parameter corresponding to the temperature and D is a distance measure between two marginal distributions. In our implementation, we have used the L_p (with $p = 1$ and $p = 2$) norm distance between the vectors corresponding to the marginal distributions; other distances can also be used. Under this probability model, the problem of generating images from $\Omega(I^{obs})$ is to generate high probability samples under q. Following Zhu et al.,[30] we use a Gibbs sampler with annealing; other statistical sampling techniques can also be used. Fig. 3 shows four examples of synthesized textures through matching independent spectral representations. In all these examples, the independent spectral representation is first calculated from the observed image shown on the left in each panel. The two on the right in each panel are typical samples. Note that while these images in each panel have similar appearance, they are very different in the RMS distance sense.

A direct application of the above sampling procedure to synthesizing objects with global topological structures is not satisfactory as the translation greatly affects the appearance of the synthesized object. To overcome this problem, we propose to use some objects as the boundary condition for sampling in that those pixels are fixed and are not updated. Fig. 4 shows an example. Fig. 4(a) shows the observed image of a telephone and Fig. 4(b) shows the image with the boundary condition. Fig. 4(c) shows a typical initial condition for sampling, which is a white noise image. Fig. 4(d)-(f) show three typical synthesized images by running the Gibbs sampler three times with different initial conditions. We can see that all the synthesized images capture the topological structures of the telephone object even though there are considerable variations in pixel values. The effectiveness of the independent spectral representation is due to the local constraints from different filters that are sensitive to local patterns and the global constraints imposed by the marginals. Note that the shapes of filters are critical. For comparison, Fig. 4(g) shows a synthesized example using principal filters, i.e, the principal components estimated from 40,000 image windows of different sizes. Even though the Gibbs sampler remains effective, the synthesized image does not capture the topological structures. Figure 5 shows three more object synthesis examples. In all the examples, the global topological structures and local structures are effectively captured by their independent spectral representation.

4. Recognition Experiments

To demonstrate the advantages of the proposed representation for recognition, we have applied it to texture classification and face recognition, representing two typical recognition problems in computer vision. For each dataset, we first estimate the independent filters using the FastICA algorithm on 40,000 samples with different window sizes randomly picked from the training images. We use window sizes from 3×3 to 25×25 and result in a total of 483 filters. We have used the rows of estimated W's as filters. (We have experimented with filters given by W_o and obtained similar results.) We then choose 40 filters using a proposed filter selection algorithm (see below). Each image is then represented by its independent spectral representation using the 40 chosen filters. The classification is done using a standard multiple layer perceptron neural network trained with back propagation;[34] we have used the nearest neighbor rule and other classifiers and in most cases we have obtained comparable results.

In this section, we first present our filter selection algorithm and then our recognition results on a texture dataset, a face dataset, and two infrared datasets. We have included the infrared datasets to show that the proposed representation is generalizable to sensors other than visible ones.

A. Filter Selection Algorithm

As we discussed earlier, the FastICA algorithm estimates the independent filters at a particular window size. By applying the algorithm with different window sizes, we obtain a large number of independent filters. In this paper, we have in total 483 filters at scales from 3×3 to 25×25. These filters however are redundant and not necessary for a particular dataset. The redundancy is due to that the marginal distribution of two different filters can be identical if the filters are identical after a translation. Filters at different scales may also have similar marginal distributions as they are estimated separately. In addition, the large number of filters is computationally expensive as the calculation of the independent spectral representation requires convolution with different filters.

To select a small number of filters for recognition, here we use a filter selection algorithm that was proposed by Liu and Wang for texture classification.[32] We have extended this algorithm by bootstrapping the design samples into multiple sets of training and validation sets. The basic idea is to choose filters one by one so that the performance estimated on the validation set is maximized. More formally, we divide all the available design samples into L training-validation pairs, $T^{(1)}, V^{(1)}, \ldots, T^{(L)}, V^{(L)}$, where $T^{(i)} \cup V^{(i)}$ is the full set, and $T^{(i)} \cap V^{(i)} = \phi$, for $i = 1, \ldots, L$. Then we estimate the recognition performance on $V^{(i)}$ using the training samples in $T^{(i)}$ and the filters in C, denoted by $F(V^{(i)}|T^{(i)}, C)$. The average performance is taken as an estimate of the recognition performance of the filters in C. We choose the filter that works the best with the currently chosen ones. This procedure is repeated until the performance does not increase significantly by adding one more filter. This greedy algorithm is computationally efficient and our results indicate the selected filters are sufficient for our recognition experiments even though the chosen filters are not guaranteed to be the optimal set. The algorithm is outlined below, where W is the set of the candidate independent filters estimated from the given dataset, C the subset of filters that has been chosen so far, L a parameter that specifies the number of training-validation pairs, and ϵ a threshold.

Filter Selection Algorithm

$W = \{\vec{W}_1, \cdots, \vec{W}_K\}, \quad C = \phi$

Generate randomly L training-validation pairs $T^{(1)}, V^{(1)}, \ldots, T^{(L)}, V^{(L)}$.

repeat

for each filter $\vec{W}_\alpha$ in W

calculate $F(C \cup \vec{W}_\alpha) = 1/L \sum_{i=1}^{L} F(V^{(i)}|T^{(i)}, C \cup \vec{W}_\alpha)$

$\alpha^* = \max_\alpha F(C \cup \vec{W}_\alpha)$

$E = F(C \cup \vec{W}_{\alpha^*}) - F(C)$

if $E > \epsilon$ then

$C = C \cup \vec{W}_{\alpha^*}, \quad W = W \setminus \vec{W}_{\alpha^*}$

until $E \leq \epsilon$

B. Texture Classification

We have applied our method to the problem of texture classification, which has been studied extensively as a separate topic in computer vision. We argue that texture models should be consistent with perceptual models for objects as they need to be addressed within one generic recognition system; we demonstrate in this paper that our method can be applied equally well to texture classification as well as to face and object recognition.

To demonstrate the effectiveness of our approach, we use a dataset consisting of 40 textures, as shown in Fig. 6. Each texture image is partitioned into non-overlapping patches with size 32×32 and then all the obtained patches are divided into a training set and a test set with no common patch between the two sets. We first estimate the independent filters using the FastICA algorithm on 40,000 samples randomly picked from the training set. We then apply our filter selection algorithm to select 40 filters from all the independent filters. The neural network trained with the chosen filters is then used to classify the patches in the test set. To avoid the bias due to the choice of the training set, we randomly choose the training set for each texture and run our algorithm many times for a better evaluation. We also change the number of patches in the training set to demonstrate the generalization capability of our representation.

The 40 selected filters are shown in Fig. 7. We can see that our algorithm chose filters whose shape is comparable with dominant local texture patterns. Table 1 shows the classification result with 100 trials for each setting. This dataset is very challenging in that some of textures are perceptually similar to other textures in the dataset and some are inhomogeneous with significant variations. With as few as 8 training patches, our method achieves a correct classification rate of 92% on average. With half patches used for training, we achieve an average classification rate over 96%.

For comparison, Table 1 also shows the results using principal and independent component representations directly on each 32×32 image patch. Each training patch is represented by a vector and we then train a classifier using the vectors in the training set. The trained classifier is then used to classify test patches also represented by a vector. As discussed earlier, clearly the direct use of linear representations is not effective for texture classification.

To further demonstrate the effectiveness of the proposed representation, we have applied our method to two challenging texture groups for classification used by Randen and Husoy[35] in a comprehensive comparative study. The two groups, shown in Fig. 8, are difficult due to the significant variations and inhomogeneity within each texture type and the similarities between some textures. These factors result in large within-class distances and possible small between-class distances, making the classification of these groups very challenging. The average recognition rate of all the methods studied in Randen and Husoy[35] is 52.6% for Fig. 8(a) and 54.0% for Fig. 8(b) with the best recognition rate of 67.7% and 72.2% respectively.

Here we use the same experimental setting as in Randen and Husoy.[35] A separate training group is used to estimate the independent filters. Then 40 filters are selected and a multiple layer perceptron is trained using 32×32 image windows from the training images. Then the trained multiple layer perceptron is used to classify 32×32 window patches in the test images. Fig. 9 shows the results from all the methods in Randen and Husoy[35] along with ours for comparison. For Fig. 8(a), our method gives a correct classification rate of 91.01% and For Fig. 8(b) 91.72%. Clearly our method gives a significant improvement over all the methods studied in Randen and Husoy.[35] This is because that the independent spectral representation is based on the underlying probability model of an image, not the image itself.

While texture classification using filters has been studied extensively, its performance is far from satisfactory partially due to that no filters are optimal for all textures and partially due to the need of more invariant texture features. As in the examples shown here, there are often considerable variations within one texture type and remarkable similarities between different textures, making the generic texture classification problem very challenging. Here we resolve the filter design problem by deriving filters as independent components of images and address the feature extraction by using the underlying probability model of the image. While more thorough experiments and analysis are to be further investigated, the results are convincing. We expect that the proposed method would provide a significant performance improvement for texture classification.

C. Face Recognition

We have also applied our method to face recognition using ORL,[36] a standard face dataset. The dataset consists of faces of 40 different subjects with 10 images each. The images were taken at different times with different lighting conditions on a dark background. While only limited side movement and tilt were allowed, there was no restriction on facial expression.

The procedure is the same as that for texture classification. We first calculate 483 independent filters using the FastICA algorithm and then select 40 from them. We vary the number of the training faces per subject and the remaining images are used for testing. As we did for texture classification, we randomly choose the training images from the dataset to avoid the potential bias due to the choice of training faces. The results are shown in Table 2. Here the variation of the performance is significant, especially the difference between the best and worst, indicating the choice of the training set can affect the recognition performance significantly. Our performance is significantly better than that of PCA and ICA essentially because different lighting conditions and facial expression make the pixel-wise based linear representation not reliable for recognition.[37] The results obtained here are also significantly better than those obtained by Zhang et al.[37] on the same dataset.

D. Infrared Face Recognition

To demonstrate that the proposed method can be generalized to images from other sensors, here we apply our method to two infrared face datasets shown Fig. 10. As in the texture and face recognition case, we first estimate the independent filters from 40,000 samples of each window size randomly picked from the training images with window sizes from 3×3 to 25×25. Then filters are selected from all the estimated independent filters.

The first dataset, partially shown in Fig. 10(a), consists of nine subjects, each with about 30 pictures under different poses and facial expressions. As oppose to video face recognition, the infrared sensor is not sensitive to lighting conditions.[38] We use roughly half of the images as training and the rest as test. We have obtained 99.05% correct recognition rate, which is considerably better than those methods reported in Srivastava and Liu[38] on the same dataset.

We have also applied our method to a large dataset, some of which are shown in Fig. 10(b). Here we use longwave infrared images of 63 subjects with a total of 3893 images. Compared to the dataset in Fig. 10(a), the variation within each class is relatively small. As in the other cases, we estimated 483 independent filters and then apply our filter selection algorithm. Because of the relative small variations within each class, our filter selection algorithm selected only one filter which gives perfect classification on all the validation sets. Thus instead of using 40 filters, only one filter is used and the performance on the test set is 99.43%.

5. Discussion

Based on the assumption that filters are statistically independent, we have proposed the independent spectral representation as feature statistics for images and demonstrated its effectiveness for recognition on several datasets. There are theoretical as well as computational issues regarding the independence assumption. Theoretically, filters that are completely independent may not exist for image datasets as images form manifolds that are intrinsically nonlinear.[39] This implies that dependence must exist in general among linear independent components, which leads to recent studies on modeling the dependence (e.g. Hyvärinen and Hoyer[40,41] and Wainwright et al.[42]). Computationally, as discussed in Sect. 2, the estimation of independent components requires the evaluation of a joint distribution in a high dimensional space; this makes it computationally infeasible to compute the globally optimal independent components given by Eq. (4). In this regard, independent component algorithms such as the FastICA algorithm[22] used here are significant in that they make it possible to compute approximate solutions of an infeasible problem. Nevertheless, none of the existing algorithms can guarantee the global optimality of its solution.

Given the theoretical and computational problems of estimating independent components, it would be important to study how the performance of the independent spectral representation relies on the independence of "independent" components. We argue that the effectiveness of the independent spectral representation can be more general. In other words, the recognition performance using the independent spectral representation can be effective even when there is significant dependence among the filters. It is obvious that the completely independent filters are preferred as they guarantee that the independent spectral representation is equivalent to the joint one. When there exists dependence among the filters, the independent spectral representation incurs an information loss. Whether this loss is critical for recognition performance depends on the training and test set. As shown by Domingos and Pazzani[43] for the naive Bayesian classifier case, this does not incur any recognition performance loss under conditions where independence is violated by a large margin. Given this, we argue that the independent spectral representation may provide a robust representation even the filters are not independent. To support this argument, we have used filters that are known to be dependent for comparison. Here we use principal components as filters. Note that the principal components here are calculated from the 40,000 samples with different window sizes from training images and they are not the principal components of entire images. Principal filters do not correlate with each other. However, if there exists higher than the second order statistical dependence, the principal components are statistically dependent. The procedure we use for recognition is essentially the same except that we use principal filters in place of independent filters. We found the performance is comparable with that of using independent components even though the independent spectral representation gives better performance in general. Table 3 gives an example using the texture dataset shown in Fig. 6.

In the spectral representation framework,[24] the filters determine the recognition performance on a particular dataset. Our method proposed here can be seen as a two-step approximate algorithm of finding optimal filters for recognition. In the first step, independent filters are learned in an unsupervised way by maximizing a criterion related to statistical independence implemented by the FastICA algorithm.[22] In the second step, optimal filters among the independent ones are selected by the filter selection algorithm. Instead of selecting filters from the independent ones, it is conceptually more appealing to learn the optimal filters by maximizing the recognition performance over all the possible filters. This, however, is a computationally challenging optimization task due to the dimension of the search space and the computational cost of estimating the performance for a given set of filters. While our empirical results have shown the independent filters provide satisfactory results on the datasets we have used, it would be interesting to compare the independent filters with the learned optimal filters for recognition. This needs to be further investigated.

6. Conclusion

In this paper we have proposed an independent spectral representation for recognition based on images. In this representation, the filters are estimated from training images as independent components of local image windows of different sizes. The sufficiency of the representation is demonstrated through texture and object synthesis experiments. For recognition on a particular dataset, a subset of filters is selected to maximize the recognition performance. We have shown empirically using different kinds of datasets its effectiveness for recognition and obtained significant improvement over commonly used linear representations. Our results support that independent components provide an effective way to overcome the curse of dimensionality by reducing the complexity of probability models and the complexity reduction does not incur a significant loss of recognition performance in the proposed representation framework.

We emphasize that our representation is proposed within the linear representation framework under the independence assumption. While our representation makes it clear that an image should be characterized by the underlying probability model and improves the recognition performance on the datasets we have used, by no means this study implies that the proposed representation or its kind is sufficient for general visual recognition. The linear representation is a crude imaging model and the estimation procedure of the underlying probability model we have used is valid only in the linear representation framework. It seems clear however that sufficient probability models not subject to the curse of dimensionality would be critical for an efficient implementation of a general vision system. We hope that our study is useful in that direction.

Acknowledgement

The authors would like to thank the reviewers whose insightful comments have significantly improved the presentation of this paper, Aapo Hyvärinen for providing several technical corrections regarding independent component analysis and references regarding the usage of independence to overcome the curse of dimensionality, and Nuno Vasconcelos and Anuj Srivastava for insightful and stimulating discussions. This research was partially supported by a National Imagery and Mapping Agency grant (NMA 201-01-1-2010) under its University Research Initiative program.

Email addresses: liux@cs.fsu.edu (X. Liu) and leicheng@cs.fsu.edu (L. Cheng).

Website links: http://www.cs.fsu.edu/∼liux (X. Liu) and http://www.cs.fsu.edu/∼leicheng (L. Cheng).

References

1. D. H. Ballard, *An Introduction to Natural Computation* (MIT Press, Cambridge, MA, 1997).

2. D. Marr, *Vision* (W. H. Freeman and Company, New York, 1982).

3. D. C. Knill and W. Richards, eds., *Perception as Bayesian Inference* (Cambridge University Press, Cambridge, UK, 1996).

4. H. von Helmholtz, *Treatise on Physiological Optics* (Dover Publications, New York, 1867).

5. R. Bellman, *Adaptive Control Processes: A Guided Tour* (Princeton University Press, New Jersey, 1961).

6. H. Hotelling, "Analysis of a complex of statistical variables in principal components," *Journal of Educational Psychology* **24**, 417–441, 498–520 (1933).

7. M. M. Loève, *Probability Theory* (Van Nostrand, Princeton, N.J., 1955).

8. L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *J. Opt. Soc. Am. A* **4**, 519–524 (1987).

9. P. Comon, "Independent component analysis, A new concept?" *Signal Processing* **36**, 287–314 (1994).

10. P. J. Huber, "Projection pursuit," *The Annals of Statistics* **13**, 435–475 (1985).

11. H. B. Barlow, "Unsupervised learning," *Neural Computation* **1**, 295–311 (1989).

12. D. J. Field, "Relations between the statistics of natural images and the response properties of cortical

cells," *J. Opt. Soc. Am. A* **4**, 2379–2394 (1987).

13. D. J. Field, "What is the goal of sensory coding?" *Neural Computation* **6**, 559–601 (1994).

14. B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature* **381**, 607–609 (1996).

15. A. J. Bell and T. J. Sejnowski, "The "independent components" of natural scenes are edge filters," *Vision Research* **37**, 3327–3338 (1997).

16. E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annual Review of Neuroscience* **24**, 1193–1216 (2001).

17. A. Srivastava, A. Lee, E. P. Simoncelli, and S. C. Zhu, "On advances in statistical modeling of natural images," *Journal of Mathematical Imaging and Vision*, in press (2003).

18. N. Vasconcelos and G. Carneiro, "What is the role of independence for visual recognition?" in *Proceedings of the 7th European Conference on Computer Vision*, vol. 1, pp. 297–311 (2002).

19. F. W. Campbell and J. G. Robson, "Application of Fourier analysis to the visibility of gratings," *J. Physiol.* (Lond.) **197**, 551–566 (1968).

20. R. L. De Valois and K. K. De Valois, *Spatial Vision* (Oxford University Press, New York, 1988).

21. A. Hyvärinen, "Survey on independent component analysis," *Neural Computing Surveys* **2**, 194–128 (1999).

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

23. R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. (Wiley, New York, 2000).

24. X. Liu and D. L. Wang, "A spectral histogram model for texton modeling and texture discrimination," *Vision Research* **42**, 2617–2634 (2002).

25. D. A. Socolinsky and A. Selinger, "A comparative analysis of face recognition performance with visible and thermal infrared imagery," in *Proceedings of the International Conference on Pattern Recognition*, vol. 4, pp. 217–222 (2002).

26. J. R. Bergen and E.H. Adelson, "Early vision and texture perception," *Nature* **333**, 363–367 (1988).

27. C. Chubb, J. Econopouly, and M. S. Landy, "Histogram contrast analysis and the visual segregation of IID textures," *J. Opt. Soc. Am. A* **11**, 2350–2374 (1994).

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

29. S. C. Zhu, Y. N. Wu, and D. Mumford, "Minimax entropy principle and its application to texture modeling," *Neural Comp.* **9**, 1627–1660 (1997).

30. S. C. Zhu, X. Liu, and Y. N. Wu, "Exploring texture ensembles by efficient Markov Chain Monte Carlo," *IEEE Trans. Pattern Anal. Machine Intell.* **22**, 554–569 (2000).

31. X. Liu and A. Srivastava, "3D object recognition using perceptual components," in *Proceedings of the*

International Joint Conference on Neural Networks, vol. 1, pp. 553–558 (2001).

32. X. 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).

33. X. Liu, D. L. Wang, and A. Srivastava, "Image segmentation using local spectral histograms," in *Proceedings of the International Conference on Image Processing*, vol. 1, pp. 70–73 (2001).

34. J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Reading, MA, 1991).

35. T. Randen and J. H. Husoy, "Filtering for texture classification: A comparative study," *IEEE Transactions on Pattern Recognition and Machine Intelligence* **21**,291–310 (1999).

36. Images in the ORL dataset are available at http://www.uk.research.att.com/facedatabase.html.

37. J. Zhang, Y. Yan, and M. Lades, "Face recognition: Eigenface, elastic matching, and neural nets," *Proceedings of IEEE* **85**, 1423–1435 (1997).

38. A. Srivastava and X. Liu, "Statistical hypothesis pruning for identifying faces from infrared images," *Image and Vision Computing*, in press (2003).

39. S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science* **290**, 2323–2326 (2000).

40. A. Hyvärinen and P. O. Hoyer, "Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces," *Neural Computation* **12**, 1705–1720 (2000).

41. A. Hyvärinen and P. O. Hoyer, "A two-layer sparse coding model learns simple and complex cell receptive fields and topology from natural images," *Vision Research* **41**, 2413–2423 (2001).

42. M. J. Wainwright, E. Simoncelli, and A. S. Willsky, "Random cascades on wavelet trees and their use in analyzing and modeling natural images," *Applied and Computational Harmonic Analysis* **11**, 89–123 (2001).

43. P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," *Machine Learning* **29**, 103–130 (1997).

List of Figures

Fig. 1. Independent filters computed from the 40-texture dataset shown in Fig. 6. In each panel, each image in the left corresponds to one column of $\mathbf{S}$, in the middle corresponds to one row of $\mathbf{W}$, and in the right corresponds to one row of $\mathbf{W_o}$. (a)-(c) Independent filters of scales 5×5, 7×7, and 11×11 respectively.

Fig. 2. The number of required cells to represent a joint distribution (solid line) and the corresponding distributions discretely with respect to the dimensionality, assuming each dimension is quantized into 10 bins. The vertical axis is shown in logarithmic scale.

Fig. 3. Synthesized textures through matching independent spectral representations. In each panel, the left shows the original image and the remaining two show two typical examples of synthesized images. (a) A texture with rough periodic elements. (b) A texture with stochastic horizontal elements. (c) A texture with no obvious repeated patterns. (d) A texture with detailed elements.

Fig. 4. Synthesized images of an object. (a) A telephone object given as an image. (b) The input with the boundary condition used for synthesis. (c) Initial condition for synthesis, which is a white noise image. (d)-(f) Three synthesized images of the object through matching independent spectral representation. (g) A synthesized image using principal filters.

Fig. 5. Synthesized images of objects and faces. In each panel, the left most is the input image and the rest are typical synthesized examples by matching independent spectral representations. Here similar objects are used as boundary conditions as in Fig. 4 but are not shown here. (a) An stapler. (b) and (c) Two faces.

Fig. 6. Forty natural textures used in the classification experiments (available at http://www-dbv.cs.uni-bonn.de/image/texture.tar.gz). The input image size is 256×256.

Fig. 7. 40 filters selected for the texture dataset shown in Fig. 6.

Fig. 8. 10-texture image groups used in Randen and Husoy.[35] Each image is 128×128. (a) The texture images in Figure 11(h) of Randen and Husoy.[35] (b) The texture images in Figure 11(i) of Randen and Husoy.[35]

Fig. 9. Correct classification rate for all the methods in Randen and Husoy[35] for Fig. 8(a) and (b) respectively and the proposed method. In each plot, each data point represents one result (corresponding to one texture classification method) in Tables 3, 6, 8, and 9 of Randen and Husoy,[35] and the dashed line is the result of the proposed method.

Fig. 10. (a) Example images from the FSU IR face database (available at http://fsvision.fsu.edu). (b) Long wave infrared image examples from a dataset of 3893 images generated by Equinox company (available at http://www.equinoxsensors.com/products/HID.html.

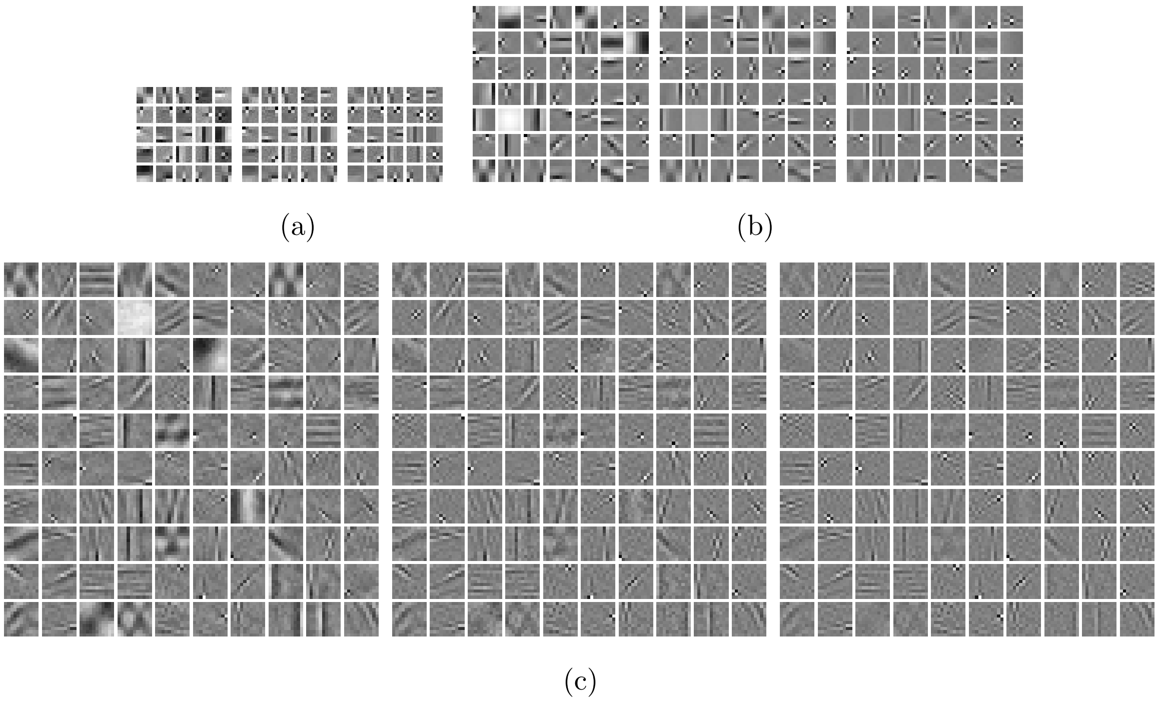


Figure 1, Liu and Cheng

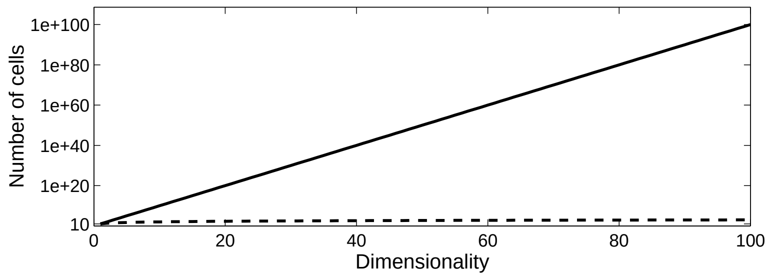
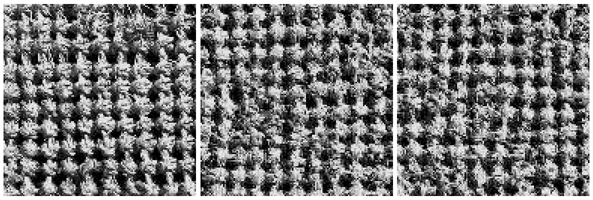
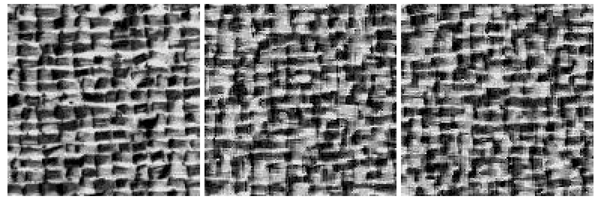


Figure 2, Liu and Cheng

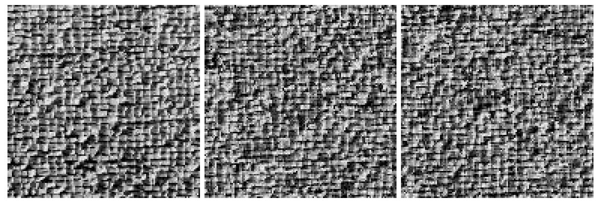


(a)



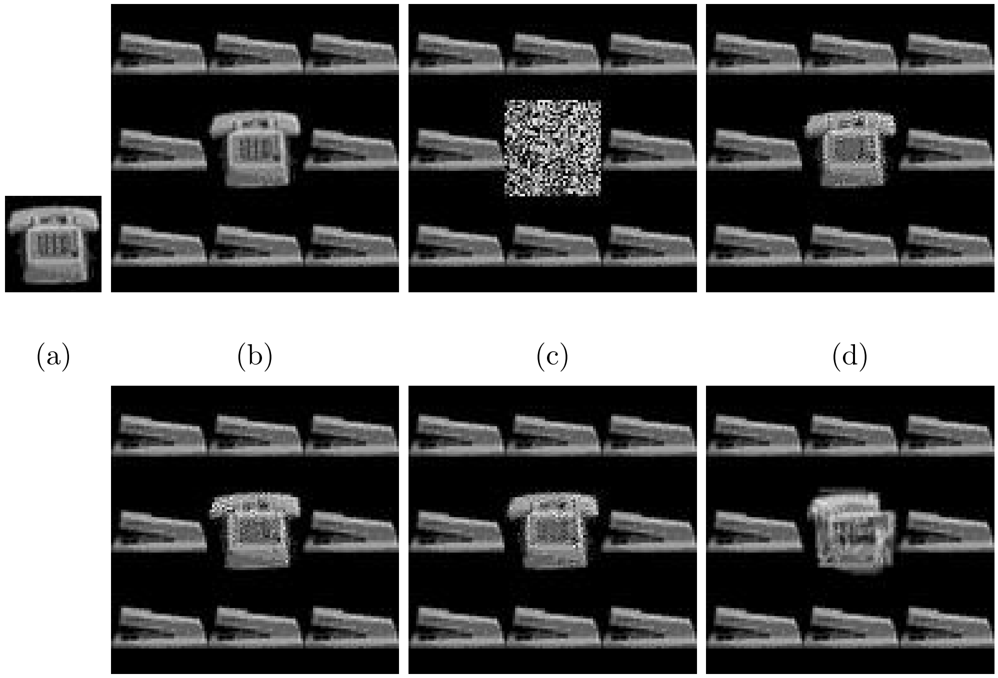
(b)

(c)



(d)

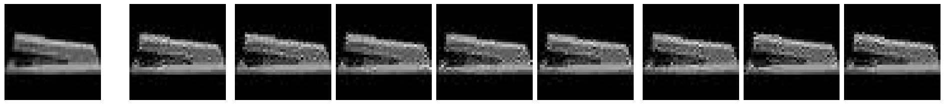
Figure 3, Liu and Cheng



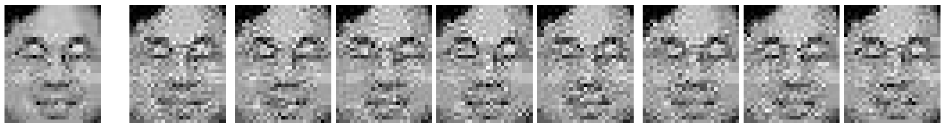
(a) (b) (c) (d)

(e) (f) (g)

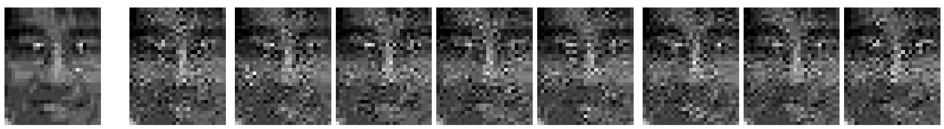
Figure 4, Liu and Cheng



(*a*)



(*b*)



(*c*)

Figure 5, Liu and Cheng

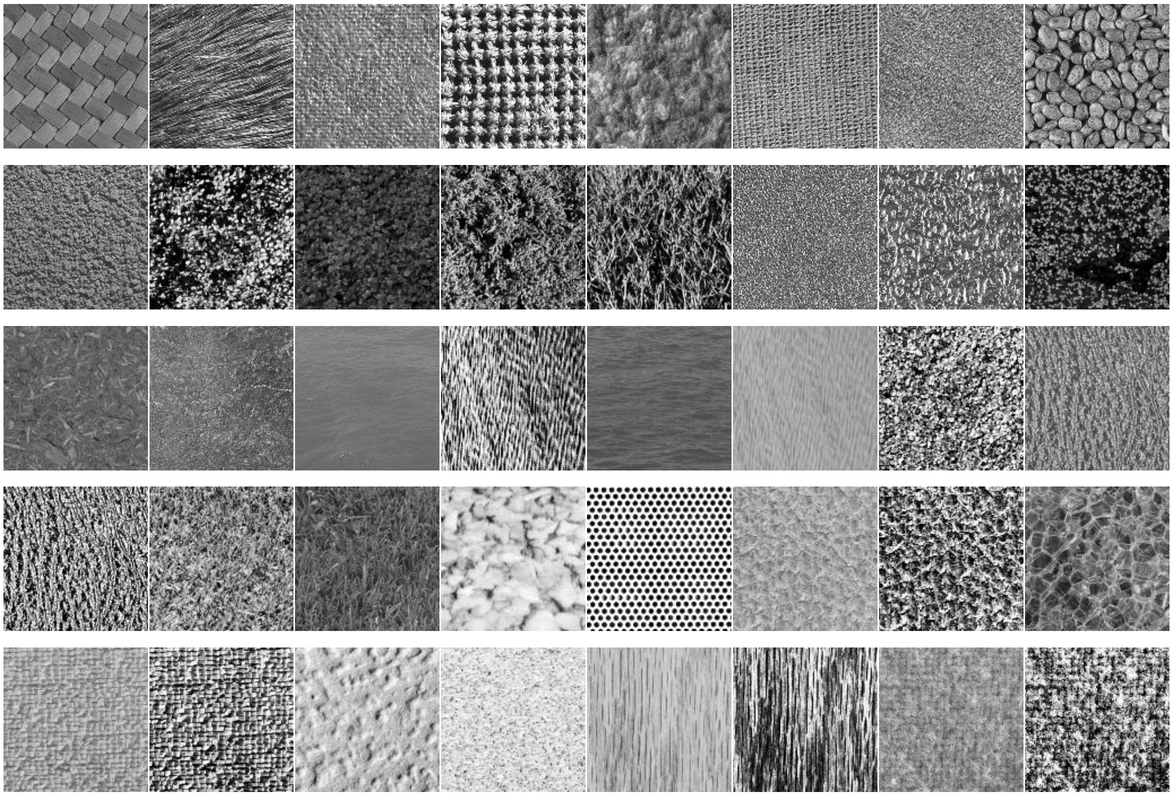


Figure 6, Liu and Cheng

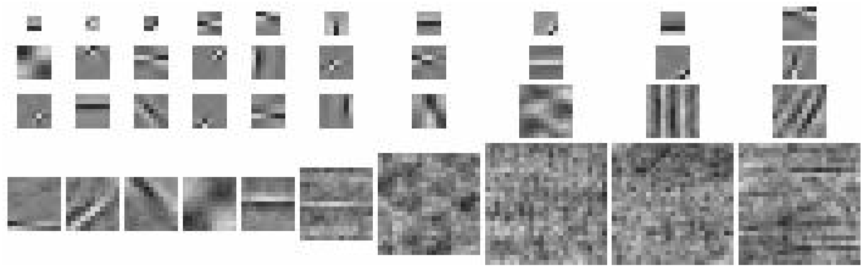
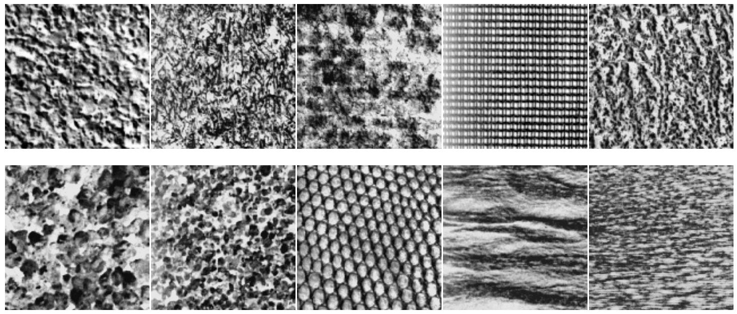
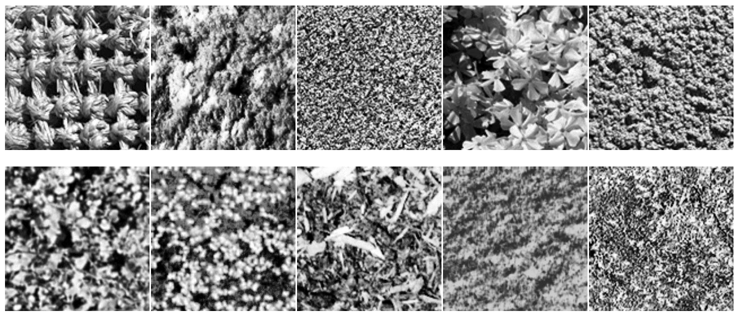


Figure 7, Liu and Cheng

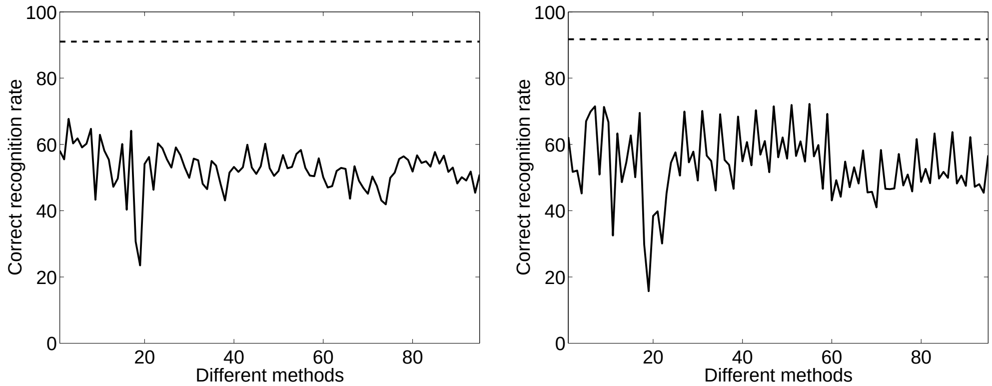


(a)



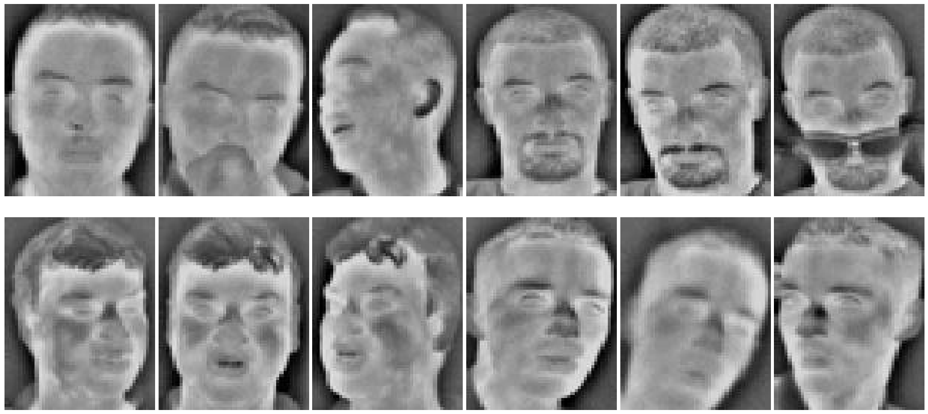
(b)

Figure 8, Liu and Cheng



(a) (b)

Figure 9, Liu and Cheng



(a)

(b)

Figure 10, Liu and Cheng

Table 1. Average recognition results for the texture dataset shown in Fig. 6

Methods	Training/test per texture			
	32 / 32	22 / 42	16 / 48	8 / 56
PCA	22.84%	20.57%	18.74%	15.88%
ICA	23.17%	21.80%	20.32%	18.54%
Proposed	97.81%	97.44%	95.42%	92.19%

Table 2. Recognition results for ORL face dataset of 100 trials

Different Methods	Training/test face images per subject					
	Average rate		Best rate		Worst rate	
	5/5	3/7	5/5	3/7	5/5	3/7
PCA	94.53%	88.00%	98.5%	92.14%	89.0%	80.71%
ICA	94.00%	85.95%	97.50%	92.50%	89.0%	75.36%
Proposed	97.88%	92.74%	100%	96.77%	95.00%	87.14%

Table 3. Average recognition results using principal and independent filters for the texture dataset shown in Fig. 6

	Training/test per texture			
Methods	32 / 32	22 / 42	16 / 48	8 / 56
Principal filters	95.44%	94.64%	94.67%	92.24%
Independent filters	97.81%	97.44%	95.42%	92.19%