ICA mixture models for image processing

Te-Won Lee
The Salk Institute, CNL
10010 N. Torrey Pines Road
La Jolla, California 92037, USA
tewon@salk.edu

Michael S. Lewicki
Carnegie Mellon University, CS & CNBC
4400 Fifth Ave., Mellon Inst., 115
Pittsburgh, PA 15213, USA
lewicki@cs.cmu.edu

Terrence J. Sejnowski
The Salk Institute, CNL
10010 N. Torrey Pines Road
La Jolla, California 92037, USA
terry@salk.edu

Abstract

We apply a probabilistic method for learning efficient image codes to the problem of unsupervised classification, segmentation and de-noising of images. The method is based on the Independent Component Analysis (ICA) mixture model proposed for unsupervised classification and automatic context switching in blind source separation [1]. In this paper, we demonstrate that this algorithm is effective in classifying complex image textures such as trees and rocks in natural scenes. The algorithm is useful for de-noising and filling in missing pixels in images with complex structures. The advantage of this model is that image codes can be learned with increasing numbers of basis function classes. Our results suggest that the ICA mixture model provides greater flexibility in modeling structure and in finding more image features than in either Gaussian mixture models or standard ICA algorithms.

1 Learning efficient codes for images

The efficient encoding of visual sensory information is an important task for image processing systems as well as for the understanding of coding principles in the visual cortex. Barlow [2] proposed that the goal of sensory is to transform the input signals such that it reduces the redundancy between the inputs. Recently, several methods have been proposed to learn image codes that utilize a set of linear basis functions. Olshausen and Field [3] used a sparseness criterion and found codes that were similar to localized and oriented receptive fields. Similar results were obtained by Bell and Sejnowski [4] and Lewicki and Olshausen [5] using the infomax ICA algorithm and a Bayesian approach respectively.

The results in this paper are along the lines of research of finding efficient codes. The main difference is the modeling of the underlying structure in mutually exclusive classes with an ICA mixture model proposed in [1]. The model is a generalization of the well-known Gaussian mixture model and assumes that the observed data in each class was generated linearly by independent components with non-Gaussian densities. The ICA mixture model uses a gradient-based expectation maximization (EM) algorithm in which the basis functions for each classes are updated using an ICA algorithm. Within each ICA class the data is transformed such that the variables are as statistically independent from each other as possible [6, 7].

In this paper, the ICA mixture model is applied to images with the goal of learning classes of basis functions capturing underlying structures of the image. The learned model can be used in many image processing applications such as image classification, segmentation, and de-noising. The results demonstrate that the ICA mixture model provides greater flexibility in modeling structure and in finding more image features than in either Gaussian mixture models or standard ICA algorithms.

2 The ICA Mixture Model

A mixture density is defined as:

$$p(\mathbf{x}_t|\Theta) = \sum_{k=1}^{K} p(\mathbf{x}_t|C_k, \theta_k)p(C_k), \tag{1}$$

where $\Theta = (\theta_1, \cdots, \theta_K)$ are the unknown parameters for the component densities $p(\mathbf{x}_t|C_k, \theta_k)$. Assume that the component densities are non-Gaussian and the data within each class are described by:

$$\mathbf{x}_t = \mathbf{A}_k \mathbf{s}_k + \mathbf{b}_k, \tag{2}$$

We use gradient ascent of the log-likelihood to estimate the parameters for each class k [1].

$$\begin{aligned} \nabla_{\theta_k} L &= \sum_{t=1}^{T} p(C_k|\mathbf{x}_t, \Theta) \frac{\nabla_{\theta_k} p(\mathbf{x}_t|\theta_k, C_k)p(C_k)}{p(\mathbf{x}_t|\theta_k, C_k)p(C_k)} \\ &= \sum_{t=1}^{T} p(C_k|\mathbf{x}_t, \Theta) \nabla_{\theta_k} \log p(\mathbf{x}_t|C_k, \theta_k). \end{aligned} \tag{3}$$

The log likelihood function in eq.3 is the log likelihood for each class. For the present model, the class log likelihood is given by the log likelihood for the standard ICA model:

$$\begin{aligned} \log p(\mathbf{x}_t|\theta_k, C_k) &= \log \frac{p(\mathbf{s}_t)}{|\det \mathbf{A}_k|} \\ &= \log p(\mathbf{A}_k^{-1}(\mathbf{x}_t - \mathbf{b}_k)) - \log|\det \mathbf{A}_k|. \end{aligned} \tag{4}$$

The adaptation is performed by using gradient ascent

$$\Delta \mathbf{A}_k \propto p(C_k|\mathbf{x}_t, \Theta) \frac{\partial}{\partial \mathbf{A}_k} \log p(\mathbf{x}_t|C_k, \theta_k). \tag{5}$$

In the basis functions adaptation, the gradient of the component density with respect to the basis functions $\mathbf{A}_k$ is weighted by $p(C_k|\mathbf{x}_t, \Theta)$. In the image processing applications the mean of the images were removed and the bias vector was set to zero. However, $\mathbf{b}_k$ can be adapted as in [1]. Because our primary interest is to learn efficient codes, we choose a Laplacian prior ($p(s) \propto \exp(-|s|)$) because it captures the sparse structure of coefficients ($\mathbf{s}_k$) for natural images. This leads to the simple infomax learning rule:

$$\begin{aligned} \Delta \mathbf{A}_k &\propto -p(C_k|\mathbf{x}_t, \Theta)\mathbf{A}_k \left[\mathbf{I} - \mathrm{sign}(\mathbf{s}_k)\mathbf{s}_k^T\right], \qquad (6) \\ \log p(\mathbf{s}_k) &\propto -\sum_n |s_{k,n}| \qquad \text{Laplacian prior} \end{aligned}$$

Equations 2, 3 and 6 are the learning rules employed for this application. The complete derivation of the learning rules for the ICA mixture model can be found in [8].

3 Unsupervised image classification and segmentation

In [1] we applied the ICA mixture model to learn two classes of basis functions for newspaper text images and images of natural scenes. The same approach can be used to identify multiple classes in a single image. The learned classes are mutually exclusive and by dividing the whole image into small image patches and classifying them we can identify a cluster of patches which encode a certain region or texture of the image. Two examples illustrate how the algorithm can identify texture in images by unsupervised classification. In the first example, four texture images were taken from the Brodatz texture dataset and put into one image. The figure 1 (a) shows the texture of four different materials: (top-left) herringbone weave, (top-right) woolen cloth, (bottom-left) calf leather and (bottom-right) raffia. Four classes of basis functions were adapted using the ICA mixture model by randomly sampling 8 by 8 pixel patches from the whole image, i.e. no label information was taken into account. One million patches were processed which took five hours on a Pentium II 400 MHz processor. The learned classes corresponded to the true classes 95% of the time. The automatic classification of the image as shown in figure 1 (b) was done by dividing the image into adjacent non-overlapping 16 by 16 pixels patches. The mis-classified patches are shown in different gray levels than the square region of the texture. On larger problems (up to 10 classes and textures), the classification error rate was not significantly different. In all experiments we used the merge and split procedure in [9] which helped to speed up convergence and avoid local minima.

Another example of unsupervised image classification using the ICA mixture model is the segmentation of natural scenes. Figure 2 (left) shows an example of a natural scene with trees and rocks. The 8 by 8 pixel patches were randomly sampled from the image and used as inputs to the ICA mixture model. Two classes of basis functions were adapted. The classification of the patches are shown in figure 2 (right). The cluster of class labels can be used to roughly segment the image into trees and rocks. Note that the segmentation may have been caused by brightness. However, very similar results were obtained on the whitened image.

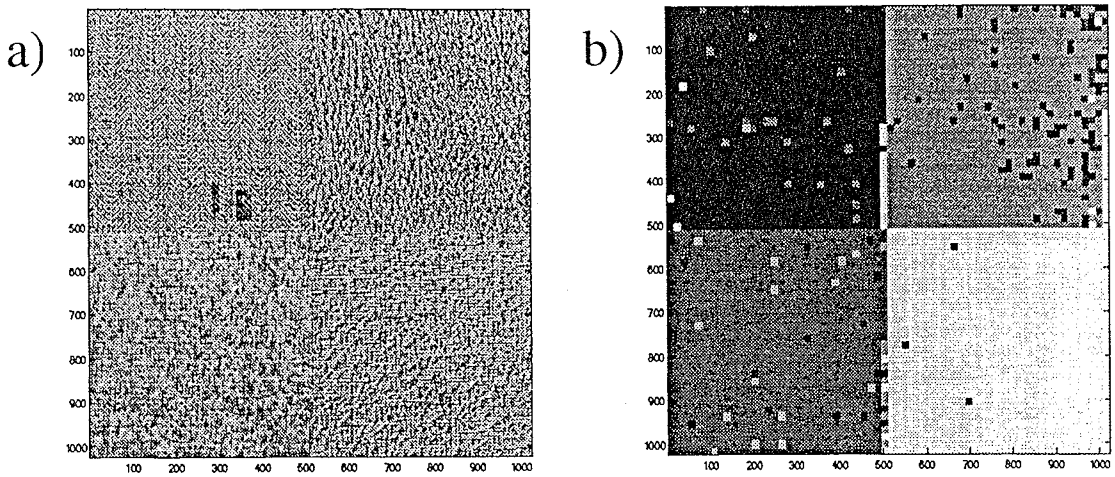

Figure 1: (a) Texture of four different materials: (top-left) herringbone weave, (top-right) woolen cloth, (bottom-left) calf leather and (bottom-right) raffia. (b) The labels found by the algorithm are shown in different gray levels. Mis-classified patches of size 16 by 16 pixels are isolated patches in a different gray level than the square region of the texture.

4 Image enhancement

The ICA mixture model provides a good framework for encoding different images types. The learned basis functions can be used for de-noising images and filling in missing pixels. Each image patch is assumed to be a linear combination of basis functions plus additive noise: $\mathbf{x}_t = \mathbf{A}_k \mathbf{s}_k + \mathbf{n}$. Our goal is to infer the class probability of the image patch as well as the coefficients $\mathbf{s}_k$ for each class that generate the image. Thus, $\mathbf{s}_k$ is inferred from $\mathbf{x}_t$ by maximizing the conditional probability density $p(\mathbf{s}_k|\mathbf{A}_k, \mathbf{x}_t)$ as shown for a single class in [5]:

$$\hat{\mathbf{s}}_k = \max_{\mathbf{s}_k} [\log p(\mathbf{x}_t|\mathbf{A}_k, \mathbf{s}_k) + \log p(\mathbf{s}_k)] \tag{7}$$

$$= \min_{\mathbf{s}_k} \left[\frac{\lambda_k}{2} |\mathbf{x}_t - \mathbf{A}_k \mathbf{s}_k|^2 + \alpha_k^T |\mathbf{s}_k| \right] . \tag{8}$$

where α_k is the width of the Laplacian p.d.f. and $\lambda_k = 1/\sigma_{k,n}^2$ is the precision of the noise for each class. The inference model in eq.8 computes the coefficients $\hat{\mathbf{s}}_k$ for each class $\mathbf{A}_k$, reconstructs the image using $\hat{\mathbf{x}}_t = \mathbf{A}_k \hat{\mathbf{s}}_k$, and computes the class probability $p(C_k|\mathbf{A}_k, \hat{\mathbf{x}}_t)$. For signal to noise ratios above 20dB the mis-classification of image patches was less than 2%. However, the error rate was higher when the noise variance was half the variance of the signal.

4.1 De-noising

To demonstrate how well the basis functions capture the structure of the data we applied the algorithm to the problem of removing noise in two different images types. In figure 3 (a) a small image was taken from a natural scene and a newspaper text. The whole image was corrupted with additive Gaussian noise that had half of the variance of the original image. The Gaussian noise changes the statistics of

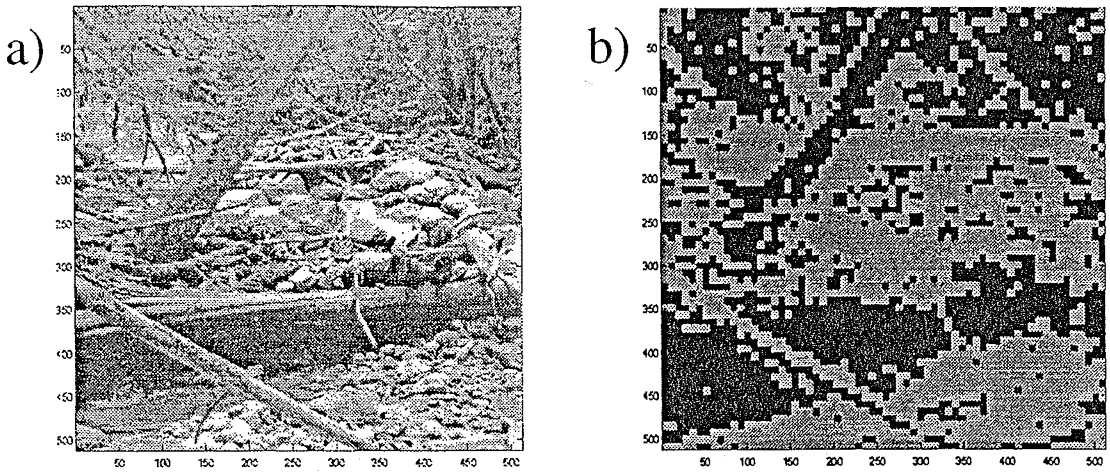

Figure 2: (Left) Example of natural scene with trees and rocks. (Right) The classification of patches (8 by 8 pixels) using the learned two sets of basis functions. The cluster of class labels can be used to roughly segment the image into trees and rocks.

the observed image such that the underlying coefficients s are less sparse than the original data. By adapting the noise level it is possible to infer the original source density by using eq.8. The adaptation using the ICA mixture model is better than the standard ICA model because the ICA mixture model is allowed to switch between different image models and therefore is more flexible in reconstructing the image. In this example, we used the two basis functions learned from natural scenes and newspaper text. For de-noising, the image was divided into small 12 by 12 pixel image patch. Each patch was first de-noised within each class and than classified by comparing the likelihood of the two classes. Figure 3 (a) shows the original image, (b) the noisy image with the signal to noise ratio (SNR) of 13dB and (c) the reconstructed image by using the Wiener filtering which a standard de-noising method with SNR=15dB. and (d) the results of the ICA mixture model (SNR=21dB). The classification error was 10%.

4.2 Filling in missing data

In some image processing applications pixel values may be missing. This problem is similar to the de-noising problem and the ICA mixture model can be used as a technique to solve this problem. In filling in missing pixels, the missing information can be viewed as another form of noise. Figure 3 (e) shows the same image with now 50% of the pixels missing. The SNR improved from 7dB to 14dB using the ICA mixture model (figure 3 (f)). The reconstruction by interpolating with splines gave SNR = 11dB. The classification error was 20%.

5 Discussion

We have investigated the application of the ICA mixture model to the problem of unsupervised classification and segmentation of images as well as de-noising, and filling-in missing pixels. Our results suggest that the method is capable of handling

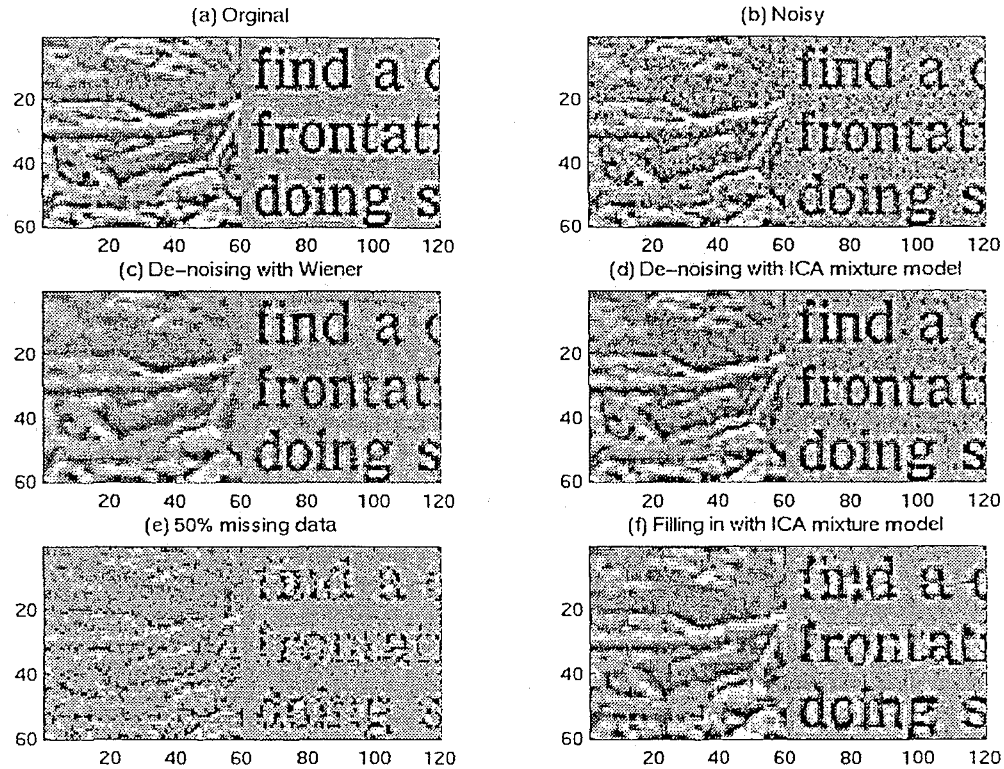

Figure 3: (a) The original image. (b) The noisy image (SNR=13dB). (c) The results of the Wiener filtering de-noising method (SNR=15dB). (d) The reconstructed image using the ICA mixture model (SNR=21dB). (e) The image with 50% missing pixels replaced with gray pixels (SNR=7dB). (f) The reconstructed image using the ICA mixture model (SNR=14dB).

the problems successfully. Furthermore, the ICA mixture model is able to increase the performance over Gaussian mixture models or standard ICA models when a variety of image types are present in the data.

The unsupervised segmentation of images by discovering image textures remains a difficult problem. Since the segmentation technique presented here is based on the classification of small image patches, the global information of the image is not taken into consideration. The multi-resolution problem may be overcome by including a multi-scale hierarchical structure into the algorithm or by re-applying the algorithm with different scales of the basis functions and combining the results. This additional process would smooth the image segmentation and the ICA mixture model could serves as a baseline segmentation algorithm. These results need to be compared with other methods, such as those proposed by De Bonet and Viola [10] which measured statistical properties of textures coded with a large-scale, fixed wavelet basis. In contrast, the approach here models image structure by adapting the basis functions themselves. The application of ICA for noise removal in images as well as filling in missing pixels will result in significant improvement when several different classes of images are present in the image. Fax machines for example transmit text as well as images. Since the basis functions of the two image models are significantly different [1] the ICA mixture model will improve in coding and enhancing the images. The technique used here for de-noising and filling-in missing pixels was proposed in [11, 5]. The same technique can be applied to multiple classes as demonstrated in this paper. The main concern of this technique is the accuracy of the coefficient prior. A different technique for de-noising using the fixed point ICA algorithm was proposed in [12] which may be intuitively sound but requires some tweaking of the parameters.

Another issue not addressed in this paper is the relevance of the learned codes to neuroscience. The principle of redundancy reduction for neural codes is preserved by this model and some properties of V1 receptive fields are consistent with recent observations [3, 4, 5]. It is possible that the visual cortex uses overcomplete basis sets for representing images; this raises the issue of whether there are cortical mechanisms that would allow switching to occur between these bases depending on the input.

The ICA mixture model has the advantage that the basis functions of several image types can be learned simultaneously. Compared with algorithms that use one fixed set of basis functions, the results presented here are promising and may provide further insights in designing improved image processing systems.

References

[1] T-W. Lee, M. S. Lewicki, and T. J. Sejnowski. Unsupervised classification with non-gaussian mixture models using ica. In *Advances in Neural Information Processing Systems 11*, volume in press. MIT Press, 1999.

[2] H. Barlow. *Sensory Communication*, chapter Possible principles underlying the transformation of sensory messages, pages 217–234. MIT press, 1961.

[3] B. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[4] A. J. Bell and T. J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

[5] M.S. Lewicki and B. Olshausen. A probablistic framwork for the adaptation and comparison of image codes. *J. Opt.Soc., A: Optics, Image Science and Vision*, submitted, 1998.

[6] A. J. Bell and T. J. Sejnowski. An Information-Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[7] J-F. Cardoso and B. Laheld. Equivariant adaptive source separation. *IEEE Trans. on S.P.*, 45(2):434–444, 1996.

[8] T-W. Lee, M. S. Lewicki, and T. J. Sejnowski. Unsupervised classification with non-Gaussian sources and automatic context switching in blind signal separation. *IEEE Transactions on Pattern Recognition and Machine Learning*, submitted, 1999.

[9] Z. Ghahramani and S.T. Roweis. Learning nonlinear dynamical systems using an em algorithm. In *Advances in Neural Information Processing Systems 11*, 1999.

[10] J. S. De Bonet and P. Viola. A non-parametric multi-scale statistical model for natural images. In *Advances in Neural Information Processing Systems*, volume 9, 1997.

[11] M.S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Computation*, to appear, 1999.

[12] P. Hyvaerinen, A.and Hoyer and E. Oja. Sparse code shrinkage: Denoising by nonlinear maximum likelihood estimation. In *Advances in Neural Information Processing Systems 11*, 1999.