©2010 International Journal of Computer Applications (0975 – 8887)
Volume 1 – No. 11

Microarray Image Denoising using Independent Component Analysis

Arunakumari Kakumani
Department of Electronics and Communication Engineering, VNR VJIET, Bachupally, Hyderabad, Andhra Pradesh, 500090, India.

Kaustubha A. Mendhurwar
Department of Computer Science and Software Engineering, Concordia University,1455 de Maisonneuve Blvd. West, Montreal, H3G1M8, Quebec, Canada.

Rajasekhar Kakumani
Department of Electrical and Computer Engineering, Concordia University, 1455 de Maisonneuve Blvd. West, Montreal, H3G1M8, Quebec, Canada.

ABSTRACT
DNA microarrays have proved to be one of the vital breakthrough technologies for exploring the patterns of gene expression on a global scale. The differential measured gene-expression levels depend largely on the probe intensities extracted during microarray image processing. Various noises introduced during the experiment and the imaging process can drastically influence the accuracy of results. Microarray image denoising is one of the challenging pre-processing steps in microarray image analysis. In this paper, we propose denoising of microarray images using the independent component analysis (ICA). The idea of ICA *i.e.* finding the linear representation of nongaussian data so that the components are independent (or atleast as independent as possible) is exploited for denoising microarray images. Through examples, it is shown that the proposed approach is highly effective as compared to the conventional discrete wavelet transform and statistical methods.

Keywords
Denoising, independent component analysis, microarray image, shrinkage function, white Gaussian noise.

1. INTRODUCTION

Microarray technology has created a great impact in bioinformatics over the last decade by its unprecedented capacity to monitor the expression levels of thousands of genes simultaneously. This technique has evolved multifold from its introduction [22] and has now become a high-throughput technology to simultaneously measure ribonucleic acid (RNA) abundances of tens of thousands of messenger ribonucleic acids (mRNA).

Microarray experimental procedure involves a number of error-prone steps which introduce high noise in the resulting data. Microarray images when corrupted with noise may drastically affect the resulting gene-expression profile. Hence, denoising of microarray image is a challenging task in the preprocessing step of microarray data analysis. Traditionally, the noise introduced is estimated using statistical methods, which include analysis of variance by Kerr [16], ratio distribution by Chen [3] and Ermolaeva [8], Gamma distribution by Newton [19], empirical Bayes model by Lonnstedt and Speed [17], and Bayesian Estimation of Array Measurements (BEAM) by Dror [7]. These methods mainly estimate the measurement error, such as preparation of the sample, cross hybridization, and fluctuation of fluorescence value from gene to gene. Recently, thresholding techniques [5] for image enhancement have become more popular

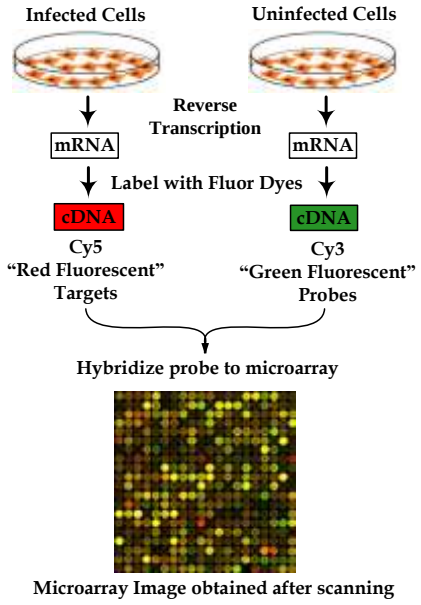


Figure 1. Schematic diagram of microarray experiment.

for eliminating such noise and ensure better gene-expression. Unfortunately, the performance of most of these methods depends on various factors such as, the type of soft/hard thresholding [6] used (*e.g.* OracleShrink, VisuShrink, SureShrink, BayesShrink), the correct estimation and fine adjustment of the threshold parameter. Moreover, the thresholding in wavelet domain needs to be applied for each of the several decomposition levels. Hence, a new method without the above mentioned constraints and which depends exclusively on the image characteristics is in demand.

In this paper, we propose an approach based on the independent component analysis [4][15] to eliminate the inherent noise in microarray images. ICA is a statistical technique in which the observed random data are linearly transformed into components that are maximally independent from each other.

2. MICROARRAY EXPERIMENT

Microarrays are grids of thousands of different single-stranded DNA molecules immobilized on a surface to serve as probes. Two major kinds of arrays are those using synthesized oligonucleotides

and those arrays using spotted cDNAs (complementary-DNA molecules). The basic procedure is shown in Figure 1, where RNA is first extracted from cells and converted to single-stranded cDNA. Then florescent labels are attached to the different cDNAs, allowing the single stranded cDNAs to hybridize to their complimentary probes on the microarray. The resulting fluor-tagged hybrids are detected *via* excitation of the attached fluors and image formation using a scanning confocal microscope. Relative RNA abundance is measured *via* measurement of signal intensity from the attached fluors. This intensity is obtained by image processing. The resulting microarray image intensities are subject to statistical analysis with particular attention to the detection of high or low expressing genes, expression based phenotype classification and the discovery of multivariate inter-gene predictive relationships. Thus, microarrays have revolutionized molecular biology research and genomic clinical diagnosis.

Since the florescent tags (*e.g.*, the red fluorescent dye Cy5 and the green fluorescent dye Cy3), are attached to the cDNA strands that hybridize, the corresponding spots on the array will fluoresce when provided fluorescence excitation energy and be detected at the level of emitted light. This yields a 16-bit tagged digital image corresponding to each colour and the intensities of the combined image reflect levels of measured fluorescence, which in turn reflect mRNA abundances. Figure 1 also shows an example of the microarray image. Microarray images are usually not perfect and are corrupted with noise that interferes in the measurement of gene-expression levels. The noise in the images originates from different sources during the course of experiment, such as photon noise, electronic noise, laser light reflection, dust on the slide, and so on. Hence, it is crucial to denoise the resultant image for accurate gene-expression profiling.

Precautionary methods to reduce the noise source include using clean glass slide and using a higher laser power rather than higher PMT voltages. However, these are not adequate for the required image qualities and an enhanced software procedure embedded within the process in a much better alternative.

3. INDIPENDENT COMPONENT ANALYSIS (ICA)

ICA is similar to the blind source separation technique [1], where sources are to be found out based solely on the mixtures of the sources available. Although, this problem appears to be practically unsolvable, ICA provides solution of this problem. However, to perform ICA, one needs to have information about the sources. This puts a limitation on the use of ICA. Nevertheless, in image processing it is very useful, as all the natural images contain similar statistical information. As such, a set of noise free images can be used for the training phase of ICA. The transform obtained from ICA can then be employed for denoising any noisy image. Natural noise free images can be obtained effortlessly from any of the numerous freely available image databases such as [20].

3.1 Definition of ICA

Let $\boldsymbol{x} = [x_1, x_2, x_3, \ldots, x_m]$ be a linear mixture vector with m linear mixtures of n independent sources, $\boldsymbol{s} = [s_1, s_2, s_3, \ldots, s_n]$. The relation between mixture vector $\boldsymbol{x}$ and the source vector $\boldsymbol{s}$ can be mathematically expressed as

$$x_j = A_{j1}s_1 + A_{j2}s_2 + A_{j3}s_3 + \ldots + A_{jn}s_n, \quad \text{for } j = 1, 2, \ldots, m, \qquad (1)$$

where $\boldsymbol{A}$ is called the mixing matrix of size $(m \times n)$, and each

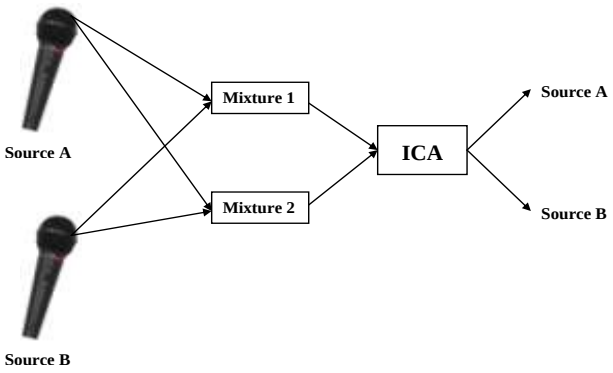


Figure 2. Block diagram of ICA for two sources is shown.

column $A_1, A_2, \ldots, A_n$ is called as basis function. As such, the basis functions project the independent sources to produce the linear mixtures. Thus, (1) can be expressed mathematically as

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}. \qquad (2)$$

The above statistical model in (2) is called as independent component analysis or the ICA model [23]. Figure 2 depicts the block diagram of ICA where $\boldsymbol{s}$ has two elements. Mixtures obtained from $\boldsymbol{s}$ are fed to ICA as inputs, and the objective is to find the individual sources. The independent/source components $\boldsymbol{s}$ are also called '*latent variables*', as they are not directly observed. Mixing matrix $\boldsymbol{A}$ is also not known. Only, mixture vector $\boldsymbol{x}$ is available, from which $\boldsymbol{A}$ and $\boldsymbol{s}$ are to be estimated.

ICA starts with a simple assumption that elements of vector $\boldsymbol{s}$ are statistically independent. The distributions of elements of $\boldsymbol{s}$ although unknown, are assumed to be *non-gaussian*, as ICA works only for *non-gaussian* distributions. Furthermore, for the sake of simplicity, mixing matrix $\boldsymbol{A}$ is assumed to be a square matrix. The objective of ICA is to find a matrix $\boldsymbol{W}$, which when multiplied with observed mixture vector $\boldsymbol{x}$, should output the source vector $\boldsymbol{s}$ *i.e.*

$$\boldsymbol{s} = \boldsymbol{W}\boldsymbol{x}. \qquad (3)$$

$\boldsymbol{W}$ is referred to as separation matrix and its inverse results in mixing matrix $\boldsymbol{A}$. Thus, ICA results in individual elements or sources of $\boldsymbol{s}$. However, the sources need not be in the same order. There are numerous ways of finding matrix $\boldsymbol{W}$, but in this work *fixed- point FastICA algorithm* [9] is used.

3.2 Fixed-Point FastICA

There are numerous ways of finding matrix $\boldsymbol{W}$. Bell and Sejnowski have proposed a neural learning algorithm based on gradient based approach [1][2]. However, gradient based methods have a drawback, that the type of the sources to be found out should be known. In general, the gradient based methods have poor convergence properties. As a remedy to this, A. Hyvärinen and E. Oja have proposed a new fast-point algorithm [11][15]. This algorithm is employed in this work and is briefly discussed below.

Sample data obtained from noise free images, say $\boldsymbol{x}$ is first whitened and is fed to *FastICA*. The Algorithm is as follows [12]:

1. Choose an initial (random or identity) weight vector $\boldsymbol{w}$.
2. Let $\boldsymbol{w}^+ = E\{\boldsymbol{x}g(\boldsymbol{w}^T\boldsymbol{x})\} - E\{g'(\boldsymbol{w}^T\boldsymbol{x})\}\,\boldsymbol{w}$
3. Let $\boldsymbol{w} = \boldsymbol{w}^+ / \|\boldsymbol{w}^+\|$
4. If not converged, go back to 2.

There is a spectrum of nonlinearities or contrast functions *g* offered by *FastICA* like '*pow3*', '*tanh*', '*gauss*', '*skew*', *etc.* Suitable nonlinearity can be used depending on the nature of the problem. *Kurtosis* [23], explained later on, is applied on each vector and the vectors are ordered correspondingly. Matrix ***W*** is formed from these vectors based on the requisite of size.

4. PROPOSED APPROACH

Proposed approach works in two phases, namely, training phase, and denoising phase. Training phase is carried out only once to find the sparse matrix ***W***, and it can then be used to any number of noisy images for denoising. Both these phases, shown in Figure 3, are described below.

4.1 Training Phase

Training phase of the approach consists of three main steps, namely, database creation, ICA and kurtosis computation, and finally estimation of sparse matrix ***W***.

4.1.1 Database Creation

For the purpose of database creation, selective images, from a freely available image database [20] are taken. These images are put together to create the database as shown in Figure 3. Natural images are selected for creation of the database, since the independent components of natural scenes are edge filters [2]. Images are selected so as to include all the possible variations of statistical data.

4.1.2 ICA and kurtosis computation

Variable *ptr* is the pointer that controls the row number of the sparse matrix ***W***, where the new vector ***w*** is going to get added. Constant *m_size* decides the total number of rows required in matrix ***W***. In this work, value of *m_size* is set to 8 as the sample size is (8×8). Thus the computation loop terminates when *ptr* reaches 9 *i.e.* matrix ***W*** comprises of 8 rows. Within each of the computational loops, a random sample is taken from the database. *FastICA* is carried out on the sample. Kurtosis *k* is computed for ***w*** found using *FastICA*. Vector ***w*** is added to the matrix ***W*** only if the value of *k* obtained is greater than 6. Condition $k > 6$ guarantees high sparsity of the elements.

*4.1.3 Estimation of sparse matrix **W***

Orthogonalization of a matrix has several benefits. These benefits include reduction of arithmetic operations, inverse of matrix is just the transpose of it, errors are not amplified if multiplied by orthogonal matrix, magnitude of all the eigen values are 1, *etc.* Matrix ***W*** obtained is hence orthogonalized as

$$W = \text{real}\left((\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})^{-1/2}\right)\boldsymbol{W}. \quad (4)$$

Training phase terminates with the estimation of orthogonal matrix. This othogonalized matrix is then employed for denoising of noisy images.

4.2 Denoising Phase

This phase starts after completion of the training phase of the approach. Orthogonal sparse matrix ***W*** computed earlier is available at the start of this phase. This phase of the approach can also be described in three subsections, namely, the transition from spatial to ICA domain and *vice-versa*, the denoising step carried out within the ICA domain by employing the shrinkage function.

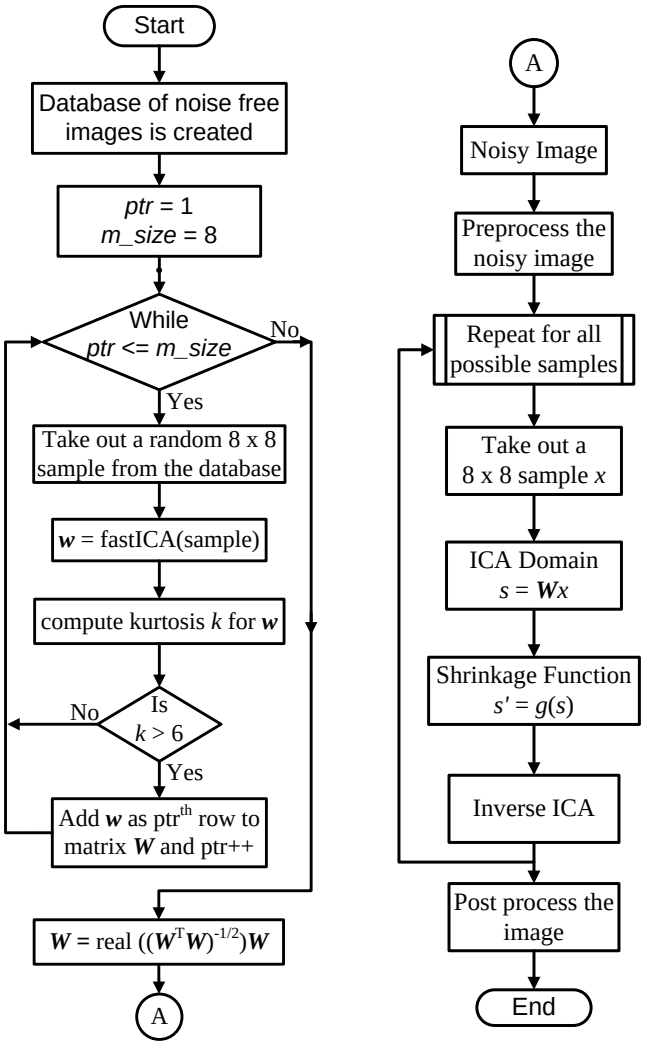


Figure 3. Flowchart of the proposed approach (a) training phase for estimating the sparse matrix, and (b) denoising in the microarray image.

4.2.1 Transition in and out of ICA domain

Sparse matrix ***W*** obtained in the training phase needs to be applied in the ICA domain. Hence, the transition in and out of the ICA domain is crucial. In the ICA domain, ***W*** can be considered as the local directional filters and coefficients in ICA domain project the image onto localized details. Transition in the ICA domain is obtained by multiplying ***W*** with the image sample. Similarly, transition out of the ICA domain is obtained by multiplying $\boldsymbol{W}^{\mathrm{T}}$ (since, ***W*** is orthogonal $\boldsymbol{W}^{-1} = \boldsymbol{W}^{\mathrm{T}}$) with the image sample.

4.2.2 Denoising

From the noisy image, recovering the original image is an impossible task; however, the aim is to estimate an image better than the noisy one. In this work, images with the additive Gaussian

noise are considered. Most of the densities encountered in image denoising can be classified as; (i) mildly sparse, and (ii) strongly sparse [14][15]. The strongly sparse density is employed in this work which for any s_i in sample $\boldsymbol{s}$ is given by [12][13]

$$p(s_i) = \frac{(\alpha+2)\ \alpha(\alpha+1)/2^{\ (\alpha/2+1)}}{\left[\alpha(\alpha+1)/2+\left|s_i/d\right|\right]^{(\alpha+3)}}. \quad (5)$$

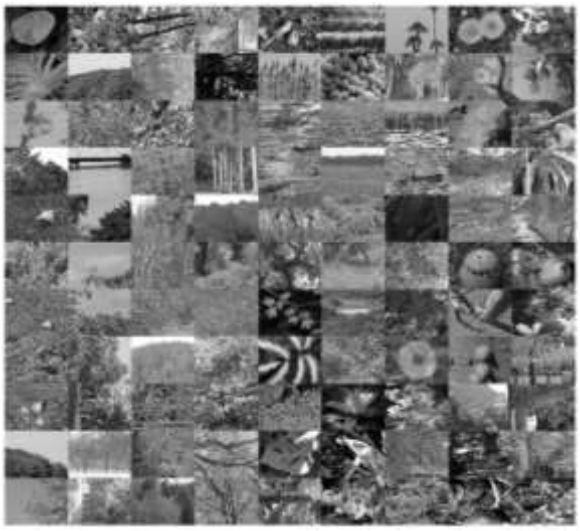


Figure 4. Database of original images is created [20].

MMSE (minimum mean square estimation) of the density model described in (6) is not easy to obtain in a closed form [15]; however, MAP (maximum a posteriori) estimation can be obtained and is given as

$$g(s_i) = \frac{1}{2d}\, sign(s_i)\, m', \quad (6)$$

where,

$$m' = \max(0, \frac{|s_i| - ad}{2} + \frac{1}{2}\sqrt{(s_i+ad)^2 - 4\sigma^2(\alpha+3)}), \quad (7)$$

$$a = \sqrt{\alpha(\alpha+1)/2}, \quad (8)$$

$$\alpha = \frac{2-k+\sqrt{k(k+4)}}{2k-1}, \quad (10)$$

$$k = d^2 P_s\ \ 0\ ^2. \quad (11)$$

The value of $P_s(0)$ is approximated to be 0.707. Shrinkage function g employed in the ICA domain is given in (6).

In essence, denoising phase of the proposed approach can be summarized as follows. A sample is taken from the image and is projected in ICA domain using $\boldsymbol{W}$. The shrinkage function g is applied on the sample, and the sample is taken out of the ICA domain using $\boldsymbol{W}^{\mathrm{T}}$. The process is repeated for all possible image samples of the noisy image.

5. RESULTS

The proposed approach is applied to microarray images and is also compared with the existing denoising approaches. The comparison results are presented in Figures 6 and 7 and are tabulated in tables 1 and 2 respectively. Following are the assessment parameters [18]

(a) (b)

Figure 5. (a) Basis functions, and (b) orthogonalized basis functions.

used to evaluate the performance of noise reduction.

A. Absolute Average Difference (AAD):

$$ADD = \frac{\sum_{r,c} |I(r,c) - I_d(r,c)|}{R \times C} \quad (12)$$

B. Signal to Noise Ratio (SNR):

$$SNR = \frac{\sum_{r,c} I(r,c)^2}{\sum_{r,c} (I(r,c) - I_d(r,c))^2} \quad (13)$$

C. Peak Signal to Noise Ratio (PSNR):

$$PSNR = \frac{R \times C \max_{r,c}(I(r,c)^2)}{\sum_{r,c} (I(r,c) - I_d(r,c))^2} \quad (14)$$

D. Image Fidelity (IF):

$$IF = 1 - \frac{1}{SNR} \quad (15)$$

E. Correlation Quality (CQ):

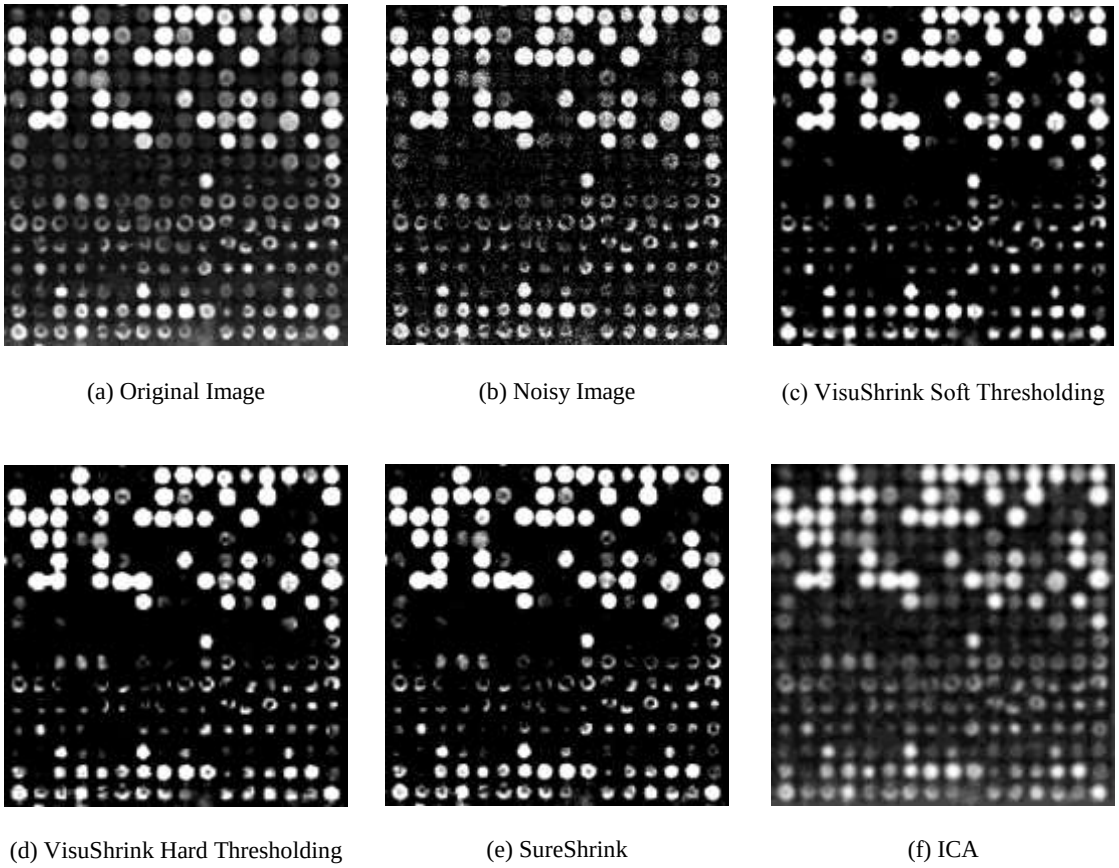
$$CQ = \frac{\sum_{r,c} I(r,c) \times I_d(r,c)}{\sum_{r,c} I(r,c)} \quad (16)$$

F. Structural Content (SC):

$$SC = \frac{\sum_{r,c} I(r,c)^2}{\sum_{r,c} I_d(r,c)^2} \quad (17)$$

where, for an image of $R \times C$ (rows-by-columns) pixels, r means row, c means column, I means original image (without noise), and I_d means denoised image. A lower *AAD* gives a "cleaner" image as more noise is reduced; larger *SNR* and *PSNR* indicates a smaller difference between the original (without noise) and denoised image; if *IF* and *SC* spread at 1, we will obtain an image I_d of better quality; and a larger value of *CQ* usually corresponds to a better quantitative performance [18].

Table 1 and Table 2 illustrate the superiority of the proposed approach as compared to the various other denoising techniques.



(a) Original Image (b) Noisy Image (c) VisuShrink Soft Thresholding

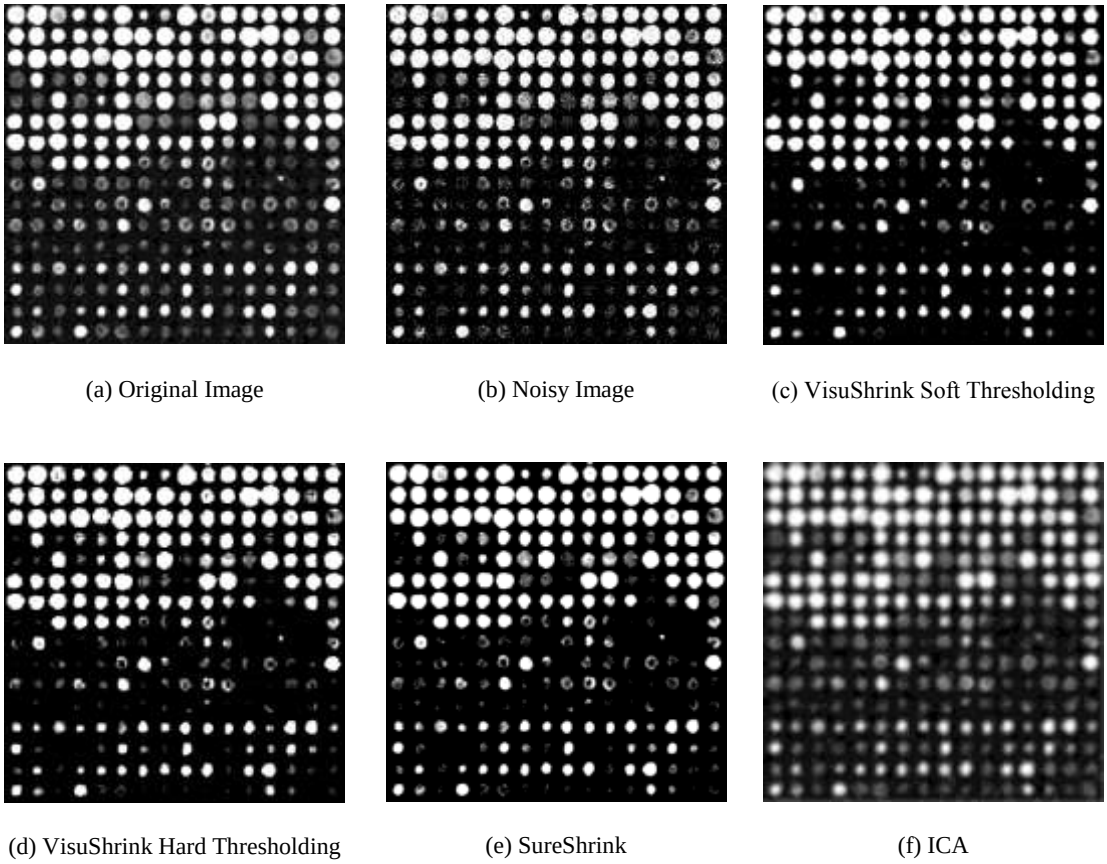
(d) VisuShrink Hard Thresholding (e) SureShrink (f) ICA

Figure 6. Original, noisy and the denoised microarray images, (a) Original Image, (b) Noisy Image, (c) VisuShrink Soft Thresholding, (d) VisuShrink Hard Thresholding, (e) SureShrink, (f) ICA.

Table 1. Values of Assessment Parameters for Various Denoising Algorithms for Figure 6(b)

Denoising Method	Assessment Parameters					
	AAD	***SNR***	***PSNR***	***IF***	***CQ***	***SC***
Mean Filter	0.3620	1.0509	56.4140	0.0484	0.4983	0.5728
Weiner Filter	0.3472	0.9350	55.9067	0.0695	0.5088	0.5259
VisuShrink (Soft)	0.3444	1.1026	56.6225	0.0930	0.4703	0.6249
VisuShrink (Hard)	0.3485	1.0828	56.5437	0.0764	0.4837	0.6005
SureShrink	0.3485	1.0844	56.5504	0.0778	0.4994	0.5813
BayesShrink	0.3491	1.0809	56.5634	0.0749	0.5006	0.5788
BivariateShrink (dwt)	0.3490	1.0951	56.5930	0.0868	0.4996	0.5840
BivariateShrink (dual tree)	0.3491	1.0977	56.6034	0.0890	0.5038	0.5797

ICA	**0.0654**	**17.2040**	**68.5447**	**0.9419**	**0.5013**	**1.1812**



(a) Original Image (b) Noisy Image (c) VisuShrink Soft Thresholding

(d) VisuShrink Hard Thresholding (e) SureShrink (f) ICA

Figure 7. Original, noisy and the denoised microarray images, (a) Original Image, (b) Noisy Image, (c) VisuShrink Soft Thresholding, (d) VisuShrink Hard Thresholding, (e) SureShrink, (f) ICA.

Table 2. Values of Assessment Parameters for Various Denoising Algorithms for Figure 7(b)

Denoising Method	**Assessment Parameters**					
	AAD	***SNR***	***PSNR***	***IF***	***CQ***	***SC***
Mean Filter	0.3642	1.1816	56.0664	0.1537	0.6123	0.5723
Weiner Filter	0.3797	1.0705	55.6381	0.0659	0.6288	0.5302
VisuShrink (Soft)	0.3649	1.1907	56.1001	0.1602	0.5752	0.6151
VisuShrink (Hard)	0.3712	1.1577	55.9780	0.1362	0.5992	0.5800
SureShrink	0.3705	1.1701	56.0245	0.1454	0.6179	0.5640
BayesShrink	0.3717	1.1616	55.9927	0.1391	0.6199	0.5600
BivariateShrink (dwt)	0.3726	1.1660	56.0092	0.1424	0.6218	0.5592
BivariateShrink (dual tree)	0.3729	1.1714	56.0289	0.1463	0.6260	0.5564
ICA	**0.0835**	**14.0653**	**66.8236**	**0.9289**	**0.6210**	**1.2345**

6. CONCLUSION

In this paper, we have proposed using independent component analysis for denoising of microarray images. The simulation results of ICA show promising results as compared to most of the popular methods such as thresholding of wavelet transform coefficients and median based filtering. ICA explicitly takes advantage of the image statistics, unlike other methods, to outperform in denoising of microarray images in terms of smoothing uniform regions, and preserving edges and features. Various assessment parameters used for evaluation of the performance of different denoising methods reveal the superiority of ICA technique. Unlike the existing denoising methods such as thresholding of wavelet transform coefficients and median based filtering, ICA explicitly takes advantage of the image statistics to outperform in denoising of microarray images. Using examples it is shown that ICA eliminates the noise without blurring edges or other sharp features of the original image.

7. REFERENCES

[1] Bell A. J. and Sejnowski T. J. 1995. An information maximization approach to blind separation and blind deconvolution. Neural Computation. Vol. 7, 1129-1159.

[2] Bell A. J., and Sejnowski T. J. 1997. The independent components of natural scene are edge filters. Vision Research. Vol. 37, 3327-3338.

[3] Chen, Y., Dougherty, E. R. and Bittner, M. L. 1997. Ratio-based decision and the quantitative analysis of cDNA microarray images. J. Biomed. Optics. 364–374.

[4] Comon P., 1994. Independent component analysis - a new concept? Signal Processing, vol. 36, 287-314.

[5] Donoho, D. L. 1995. De-noising by soft-thresholding. IEEE Trans. Inform. Theory, Vol. 41, 613–627.

[6] Donoho, D. L. 1995. Adapting to unknown smoothness via wavelet shrinkage. Journal of the American Statistical Association. Vol. 90, 1200–1224.

[7] Dror, R., Murnick, J. and Rinaldi, N. 2002. A bayesian approach to transcript estimation from gene array data: The BEAM technique. Int. Conf. on Research in Computational Molecular Biology. (Washington, USA, April 18-21, 2002) 137-143.

[8] Ermolaeva, O., et. al. 1998. Data management and analysis for gene expression arrays. Nature Genetics. Vol 20, 19–23.

[9] The FastICA MATLAB package. 1998. Available at http://www.cis.hut.fi/projects/ica/fastica/

[10] Hoyer P. 1999. Independent component Analysis in image denoising. Master's Thesis. Helsinki University of Technology.

[11] Hyvärinen A. and Oja E. 1997. A fixed-point algorithm for independent component analysis. Neural Computation, vol. 9, 1483-1492.

[12] Hyvärinen A. 1999. Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans. on Neural Networks, vol. 10, 624-634.

[13] Hyvärinen A. 1999. Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation. Neural Computation, vol. 11, 1739-1768.

[14] Hyvärinen A., *et. al.* 2000. Image denoising by sparse code shrinkage. Intelligent Signal Processing, IEEE press.

[15] Hyvärinen A., and Oja E. 2000. Independent component analysis: algorithms and applications. Neural Networks. Vol. 13, 411-430.

[16] Kerr M. K., Martin M., and Churchill G. A. 2001. Analysis of variance for gene expression microarray data. Journal of Computational Biology, vol. 7, 819–837.

[17] Lonnstedt, I. and Speed, T. 2002. Replicated microarray data. Statistica Sinica. Vol. 12, 31–46.

[18] Mastriani, M. and Giraldez, A. 2006. Microarrays denoising via smoothing of coefficients in wavelet domain. International Journal of Biomedical Sciences. Vol. 1, 7-14.

[19] Newton, M. A., et. al. 2001. On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. Journal of Computational Biology. Vol. 8, 37–52.

[20] Olmos, A. and Kingdom, F. A. A. 2004. McGill Calibrated Color Images Database. http://tabby.vision.mcgill.ca/

[21] Schena, M. 2002. Microarray Analysis. New York: John Wiley&Sons.

[22] Southern, E. M. 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. Journal of Molecular Biology. Vol. 98, 503–517.

[23] Stone, J. 2004. Independent component analysis: A Tutorial. MIT Press.