ISSN: 0976-9102 (ONLINE)
DOI: 10.21917/ijivp.2011.0022

RESTORING DISTORTED DOCUMENTS BY COMBINING NNKSOM WITH ICA AND DFDM TECHNIQUES

R. Indra Gandhi[1] and K. Iyakutti[2]

[1]*Department of Computer Applications, G.K.M College of Engineering and Technology, Tamil Nadu, India*
E-mail: shambhavi.rajesh@gmail.com
[2]*School of Physics, Madurai Kamaraj University, Tamil Nadu, India*
E-mail: iyakutti@gmail.com

Abstract

The objective of this paper is to address a special problem in character recognition from document images were the reverse side scripts appear as noise on front side and even interfere with front side characters. Due to strong background artifacts so much of double-sided distortion is noticed in ancient documents. These are often caused by the so-called bleed-through effect. Even in well-preserved documents, a similar effect called show-through is noticed because of poor paper quality. These distortions must be removed to improve readability. We propose a new NNKSOM based hybrid technique, which incorporates statistical and diffusion model to deal with bleed-through grayscale document images. The proposed method proves to perform well regardless of the intensity differences between foreground and background. This is extremely useful for researchers engaged in recognizing the distorted documents in any script worldwide as the same kind of distortion can be found in most of the scripts used in the world.

Keywords:

Distortion, Bleed-through, Show-through, Statistical and Diffusion Model

1. INTRODUCTION

Historical documents, legal documents and the like are scanned and converted to digital documents to preserve them for later use. The reproduced images might not be legible due to poor paper quality, spreading and flaking of ink, overlapping, etc. These form the basis of various kind of noise in the digitized image. There are many solutions available in the market to restore the characters from these distorted documents. But to be effective they all need clean and readable inputs. The accuracy of today's document recognition algorithms fail abruptly when document image quality distorted slightly. In addition to this, significant improvement in accuracy on hard problems now depends more on the size and quality of training sets as algorithms and hardware [1]. There are a very large number of distortions noticed in very old documents. Double-sided is a kind of distortion found generally in very old documents were a text on a side is visible on the other side, which is technically called as show-through or bleeds-through problem (Fig 1). This is one of the most challenging problems in OCR. To solve this problem we rely on two popular methods namely statistical and diffusion method (DM). But her also some disadvantage are noticed. In order to get higher degree of results for the restoration and enhancement of bleed-through documents, spatially adaptive hybrid technique is used.

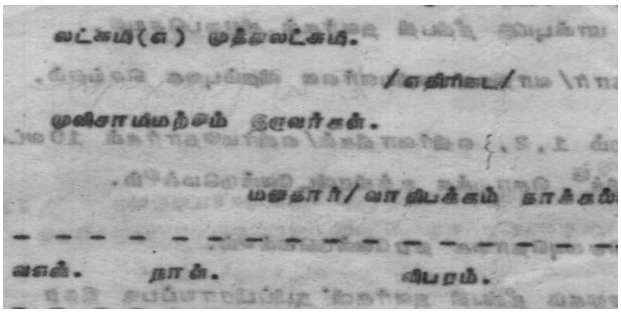


Fig.1. Double-sided distorted document

2. REVIEW OF LITERATURE

Numerous methods have been proposed in the past to recognize Bleed-though problems. In order to reach the desired goal, an ample study of research outcomes in several related areas were surveyed. Techniques of this type are reported in Knox [2] and Sharma [3] for reducing show-through in scanned documents. The basic idea is presented in [2] and a restoration technique using adaptive filtering is presented in [3]. Ophir and Malah [4] proposed a solution by taking show-through problem as a Blind Source Separation (BSS) problem, simultaneously estimating the images and mixing parameters. More over they combine a Mean Squared Error fidelity term, incorporating the non-linear mixing model and Total-Variation (TV) regularization terms applied separately to each image. Leedham et al [5] attempted the recognition process with the introduction of binarization methods with bleed through defects. Anna Tonazzini et al [6] and Emanuele et al [11] have drawn more general approaches and statistical methods such as Independent Component Analysis (ICA) and Bline Source Operation (BSS). Dubois and Anita [7] real samples are used for various distortion models. They have demonstrated the recto and the flipped verso method and using a threshold-based test to replace bleed-through with a background level. Like Dubois [7] more information regarding this can be accessed using [8]. Gang Zi [9, 10] proposed the only one other model of distortion of bleed through type of defect taking the base of blurring and mixing techniques. Xiaowei et al [12] introduced NN based approaches for show-through problem as Blind Source Separation (BSS). Moreover, there are other methods that combine several techniques such as segmentation, compression and decompression, stroke removal, etc [7, 13 and 14]. This work compares the statistical methods which are most promising with a novel approach based on the DMs. Comparison is conducted from a fundamental point of view to enable a better understanding of the advantages and disadvantages of the methods. Also, in addition to providing real samples that are obtained from [7, 8], a degradation model is developed which is capable of generating an unlimited number of document images degraded by bleed-through. This model is discussed in the next section. As known so far, there is only one other degradation model based on blurring and mixing technique [9, 10] for this type of defect. Finally, possible directions for the restoration and enhancement of very old documents are offered which benefit from the advantages of both statistical and diffusion methods.

3. METHODOLOGY

Selection of appropriate method is the common technique used to determine certain initial activities to solve the problems. Based on all those techniques, various methods were briefly introduced in this section of this paper.

3.1 STATISTICAL METHOD

Blind Signal Separation (BSS) application holds a remarkable place in statistical approach. In general BSS problem often referred as blind signal decomposition or blind source extraction (BSE) process. There appears to be something magical about blind source separation were the original source signals are estimated without knowing the parameters of mixing and/or filtering processes. It is difficult to imagine that one can estimate this. In fact, without some prior knowledge, it is not possible to uniquely estimate the original source signals. In this way the input images are considered as one-dimensional arrays, which mean that the two-dimensional input images are ignored. This is not suitable when the sources are assumed to be independent. Then the next best approach is obviously Independent Component Analysis (ICA).

3.2 INDEPENDENT COMPONENT ANALYSIS (ICA)

ICA is a newly developed statistical approach to separate unobserved, independent source variables from the observed variables that are the combinations of these source variables. Although different types of functions are used in ICA methods, the basic idea is simple. There is a cost function that determines the degree of independence of the computed sources. To obtain a best estimation, maximization of the cost function is enough and also these methods assume a linear relation between the source and the input. Using the standard ICA methodology, one can equate:

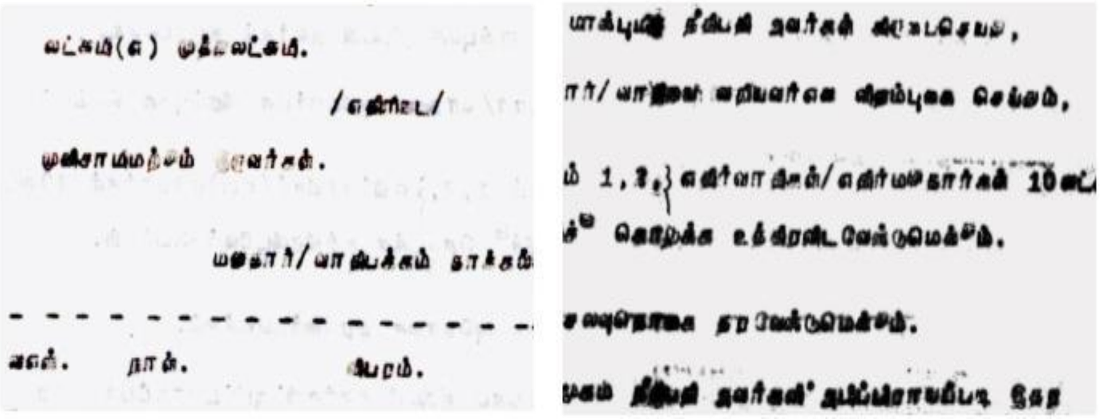
$$X = A.S \qquad (1)$$

Where X is a column matrix of mixed signals, A is a matrix representing the signal abundances and S is the column matrix of the source signals. ICA usually starts from a pre-procedure of “whitening”. The key idea here is that if the signals are independent, then they are uncorrelated, which in turn means that a procedure that de-correlates matrix X is a necessary procedure for obtaining independent signals. That is, ICA is usually performed in two stages:

$$Z = \Omega X \qquad (2)$$

$$W: \; W.Z \rightarrow \max \text{ (non-Gauss)} \qquad (3)$$

The matrix W in this case is an orthonormal matrix that can be indeed considered as a rotation matrix in the n dimensional space. The matrix Ω can be easily calculated on the basis of covariance matrix of X. Being a high-order statistical technique, ICA outperforms the second order in the discrimination power.



(a) Front Page (b) Back Page

Fig.2. Resultant Image after Applying ICA Method

3.2.1 Advantages of ICA

a) ICA usually starts from a pre-procedure of “whitening”.

b) The result of ICA processing information is near restoration to the true data.

c) It does not add any additional information other than input data.

d) Small and local fluctuations have a little effect on its output since it includes all the input data in the processing and determination of the mixing matrix.

3.2.2 Disadvantages of ICA

a) This method requires an image of two sides of the document.

b) Because of one-to one correspondence in ICA, recto and verso side of the document results will be very poor.

c) This method is so sensitive in nature. If there is any shift of co-ordination pixels due to misalignment in the scanning process. Entire result will not be clear.

d) Since this method constitute a one-dimensional signal, two-dimensional gray levels imaged are converted into one-dimensional correspondence.

e) Importantly this method does not make use of the information coming from the image nature of the inputs.

3.3 DIFFUSION METHOD (DM)

Assume that due to some distortions the true data image is destroyed and the data must be corrected via exchange of information between neighbors. These methods are based on the existence of a spatial correlation between the data of neighbouring pixels, so that each pixel is processed using the information of the surrounding pixels. This method removes all weak structures that are surrounded by the neighbouring pixels, which makes these methods very aggressive, even though it is not applicable to source separation problems. However double-sided document images can be modified to make them applicable to two-source separation problem of information (the recto and verso side). In addition to usual diffusion, some diffusion process can be added which is called double-sided flow-based diffusion method (DFDM).

3.3.1 Three ways to get Better Result

a) DFDM method cancels out the effects of real physical distortion process that occurs over time.

b) Additional diffusion processes actually separate the recto and verso side information to the background.

c) Another Reverse diffusion process is included to get better result. This not only results in uniform and fluctuation free background, but also speeds up the removal of interference by filling up the background patterns.

There are many variations of diffusion-based models available. However, the basic of all of these models is the following Eq.(4).

$$u_t = \nabla \cdot \left(c\left(\nabla_u\right)\nabla_u\right) = div\left(c\left(\nabla_u\right)\right) = DIFF(u,s,c) \qquad (4)$$

This equation can be rewritten as

$$u_t = \nabla_f \cdot \left(c_f\left(\nabla_u\right)\nabla_u\right) = div_{r,f}\left(c_{r,f}\left(\nabla_{r,u}\right)\right) = DIFF_{b,f}\left(c_{b,f}\left(\nabla_{b,u}\right)\right) \qquad (5)$$

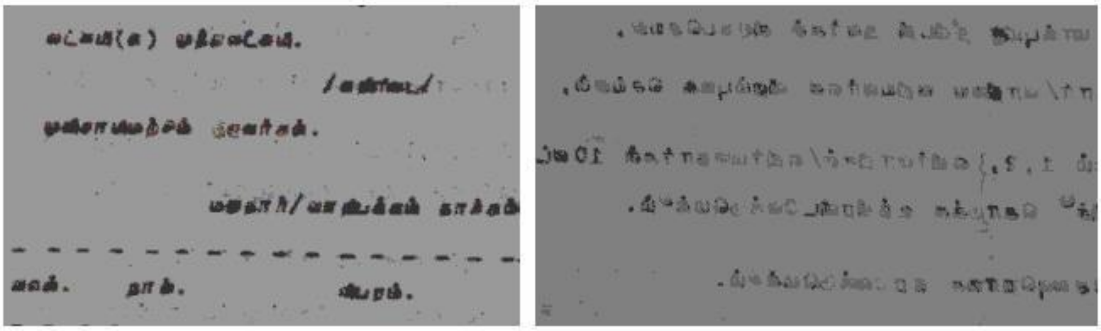
where, ‘c’ is the diffusion coefficient. Here, we introduce the extended notation of DIFF (u, s, c) to represent the diffusion process of the source ‘s’ to the target ‘u’ with the diffusion coefficient ‘c’. The ‘r’ and ‘b’ stand for the reverse and background diffusion and the flow field is denoted by ‘f’ representing a classifier from a global point of view.

3.3.2 Advantages of Diffusion Method

a) A resultant image of this method is fine and thin in structures.

b) Two-dimensional neighborhood nature collects information from the data of every pixel. Due to this, all nearby pixels will be used in the process.

c) It shows mutual local and global behavior. i.e., local behavior renders highly adaptable method to local variations same way.



(a) Front Page (b) Back Page

Fig.3. Resultant Image after Applying Diffusion Method

3.3.3 Disadvantages of Diffusion method

a) The computational cost in DM is approximately 10 times higher than ICA.

b) Sometimes it leads to negative results as the originality of the document is altered. As a result of this recognition results is low in some cases.

c) Because of restoration problems, this method is less applicable.

4. PERFORMANCE ANALYSIS

ICA has proved to be a successful technique in biomedical signal processing [15], magnetic resonance imaging analysis [16], speech recognition [17] and machine monitoring [18]. There are many more implementations of the ICA methods, such as FastICA, ICALAB [19] and Symmetric Orthogonalization[6]. As a test, we apply two different cases over Fig 1. The different cases are follows:

1. ICA Technique
2. Diffusion (DFDM) Technique

In the first case ICA method is implemented over Fig 1. The result obtained by this case is shown in Fig 2. But similar seepage of ink through a paper is noticed in Fig 2. This is a nonlinear physical phenomenon. For visualization purposes, the outputs are normalized. In the second case, diffusion method DFDM is applied over Fig 1 and the result is shown in Fig 3. The obtained result is less dominant than the interference patterns. As the interference patterns weaken, they are removed completely by the background DM. Despite the fact that the degree of bleed-through becomes so high the gray levels of the patterns are darker than those of the main text. Here DFDM diffusion method will also fail. This type of output is seen in many cases of bleed-through problems. This kind of output can be rectified by the proposed hybrid technique. In the next section, we present some hybrid technique to overcome the results and discussions seen in Fig 2 and Fig 3.

5. IMPLEMENTATION OF HYBRID TECHNIQUES

In this section, we present some combined method of ICA, DM and Neural Network (NN) based KSOM (refer Fig.4) to concentrate over Restoration and Enhancement, which a similar idea is seen in [19] but without NN.

5.1 PROPOSED HYBRID ALGORITHM

Algorithm: *Double_Sided_Restoration*

Step 1: Implement any diffusion method over input image

Step 2: Name the resulted image as DM_IMAGE

Step 3: Implement ICA method over DM_IMAGE

Step 4: Name the resulted image as ICA_IMAGE_1 and ICA_IMAGE_2

Step 5: Implement DFDM method over ICA_IMAGE_1 and ICA_IMAGE_2

Step 6: Use NN technique, to classify information for ICA and DFDM

Step 7: NN makes training for recognition

Step 8: Recognition can be done on the Content-based information

Step 9: Results restore or enhance the input document image

5.2 RESTORATION

The coefficients of the source mixture in ICA are global. Here we modify the coefficients by taking potential approach by including the results of the DM in the ICA method and include a term which computes the distance between the estimated output and the DM results.

5.3 ENHANCEMENT

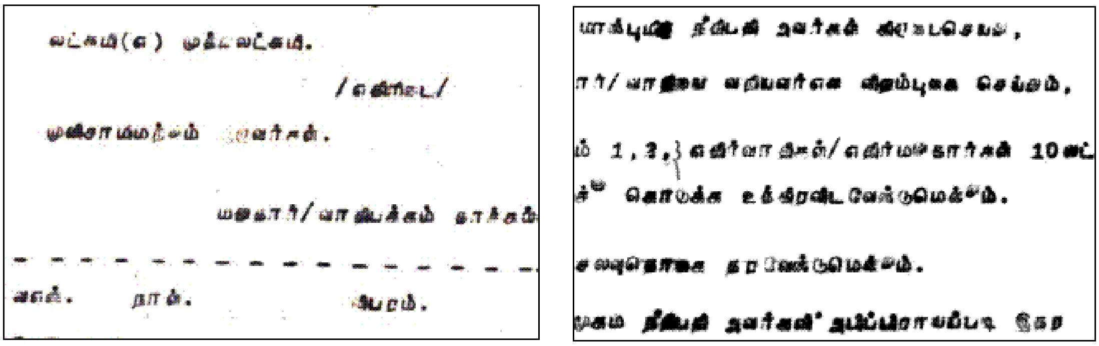
DFDM is a powerful tool for the enhancement and source separation. In general, it requires some source dominance in the inputs. Using ICA output as inputs to the DM will result in good enhancement. In this case, pre-separation using ICA will give two input images, which is very suitable for DFDM. Applying the DM then results in a very good enhancement and total separation. This type of implementation results good even in previous failed ICA methods. Still there are some defects in different colored inputs. This can be rectified using our hybrid techniques.

5.4 CONTENT BASED INFORMATION

Neural networks (NN) are richly connected networks of simple computational elements. The fundamental tenet of neural computation or computation with NN is that such networks can carry out complex cognitive and computational tasks. In addition, one of the tasks at which NN excels is the classification of input data into one of the several groups or categories. In this paper NN based KSOM is used to classify data based on content of the information (hybrid technique). The reason for using KSOM is, it is useful for visualizing low-dimensional views of high-dimensional data. It differs from the feed forward back propagations network in several ways. KSOM is trained in an unsupervised way. This means the KSOM neural network is given input data but no anticipated output. The KSOM network begins to map the training samples to each of its output neurons during training. More over KSOM does not use any sort of activation function, bias weight. Output from the KSOM does not consist of the output of several neurons it is selected as a “Winner”. Often the winning neurons represent groups in data that is presented to KSOM. Keeping all the above for the better result, we written our hybrid equations and algorithm as follows.

$$\text{DM} + \text{ICA} + \text{DFDM} + \text{NNKSOM} \qquad (6)$$

The proposed hybrid algorithm is implemented over Fig.1. The resultant output is shown in Fig. 5(a) and (b). For visualization purpose the outputs are normalized. However, all the valuable information is restored in a successful manner.



(a) Front Page (b) Back Page

Fig.5. Resultant Image after Applying Propose Hybrid Method

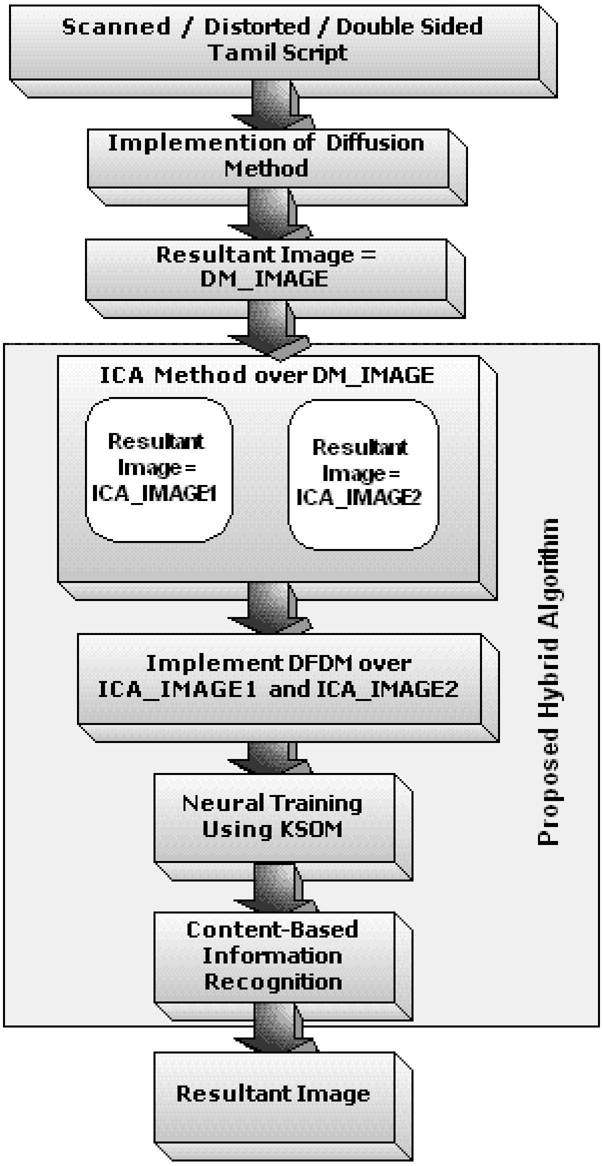

Scanned / Distorted / Double Sided Tamil Script
Implemention of Diffusion Method
Resultant Image = DM_IMAGE
ICA Method over DM_IMAGE
Resultant Image = ICA_IMAGE1
Resultant Image = ICA_IMAGE2
Implement DFDM over ICA_IMAGE1 and ICA_IMAGE2
Neural Training Using KSOM
Content-Based Information Recognition
Proposed Hybrid Algorithm
Resultant Image

Fig.4. Semantic Diagram of Proposed Hybrid Algorithm

6. CONCLUSION

In this paper we have rewritten a few formulas for DFDM. The advantages and disadvantages of ICA and DFDM method in restoring the double-sided documents are analyzed. Although ICA and DFDM produces high-resolution results in ordinary bleed-through problems of ink seepage, it is also very aggressive and seriously modifies the input data. The algorithm proposed combines these two approaches, were its efficiency being essential for applications presenting both a high degree of dimensionality and time restrictions. We therefore conclude that combining NNKSOM with ICA and DFDM restores and enhances document image more easily and results are very promising even in complex cases. A new hybrid method is introduced to gain all the advantage on both ICA and DFDM. However the hybrid method requires one more additional input. In order to fulfill this requirement ICA's two output are taken as input image for further processes.

REFERENCES

[1] Baird, H. S., 1993, “Document images defect models and their uses”, in ICDAR’93, Tsukuba, Japan, pp. 62-67.

[2] Knox, K., 1998, “Show-through correction for two-sided documents” United States Patent 5832137.

[3] Sharma, G., 2000, “Cancellation of show-through in duplex scanning”, in Proc. IEEE Int. Conf. Image Processing, Vol.2, pp.609–612.

[4] Ophir, B., and Malah, D, 2007, “Show-Through Cancellation In Scanned Images Using Blind Source Separation Techniques”, ICIP 2007, IEEE conference on Image Processing, , San Antonio, TX, Vol. 3, pp. III-233-III-236.

[5] Leedham, G., Varma, S., Patankar, A., and Govindaraju, V., 2002, “Separating text and background in degraded document images - a comparison of global thresholding techniques for multi-stage thresholding”, Proc 8th IWFHR, pp. 244–249.

[6] Anna Tonazzini, Emanuele Salerno, and Luigi Bedini, 2007, "Fast correction of bleed-through distortion in grayscale documents by a blind source separation technique," IJDAR, Vol. 10, No. 1, pp. 17–25.

[7] Dubois, E., and Pathak, A., 2001, "Reduction of bleed-through in scanned manuscript documents", in Proc. IS&T Image Processing, Image Quality, Image Capture Systems Conference (PICS2001), Montreal, Canada, pp. 177–180.

[8] Google, Book Search Dataset, Version V Edition, 2007.

[9] Gang Zi and Doermann, D., 2004, "Document image ground truth generation from electronic text", in Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th Int., Conference on, D. Doermann, Ed., Vol. 2, pp. 663–666.

[10] Gang Zi, 2005, "Ground truth generation and document image degradation", Tech. Rep. LAMP-TR-121/CARTR-1008/CS-TR-4699/UMIACS-TR-2005-08, University of Maryland, College Park.

[11] Emanuele Salerno, Anna Tonazzini, and Luigi Bedini, 2006, "Digital image analysis to enhance underwritten text in the archimedes palimpsest", IJDAR, Vol. 9, No. 2-4, pp.79–87.

[12] Xiaowei Zhang, Jianming Lu, and Takashi Yahagi, 2007, "Blind separation methods for image show-through problem," in Information Technology Applications in Biomedicine. ITAB 2007. 6th International Special Topic Conference on, Jianming Lu, Ed., pp. 255–258.

[13] Chew Lim Tan, Ruini Cao, Peiyi Shen, Qian Wang, Julia Chee, and Josephine Chang, 2000, "Removal of interfering strokes in double-sided document images," Applica-tion of Computer Vision, 5th IEEE Workshop Ruini Cao, Ed., pp.16–21.

[14] Dubois, E., and Dano, P., 2005, "Joint compression and restoration of documents with bleed - through," Proc. IS&T Archiving, Washington DC-USA, pp.170–174.

[15] Vigario, R.N., 1997, "Extraction of Ocular Artifacts from EEG using Independent Component Analysis", Electroence-phalograph. Clin. Neurophysiol, Vol. 103, pp. 395-404.

[16] Biswal, B.B., and Ulmer, J.L., 1999, "Blind Source Separation of Multiple Signal Sources of MRI Data Sets Using Independent Component Analysis," J. Comput. Assist. Tomogr., Vol. 23, pp. 265-271.

[17] Park, H. M., Jung, H.Y., Lee, T.W., and Lee, S.Y., 1999, "Subband-based Blind Signal Separation for Noisy Speech Recognition", Electronics Lett., Vol. 35, pp. 2011-2012.

[18] Ypma, A., Pajunen, P., 1999, "Rotating Machine Vibration Analysis with Second-order Independent Component Analysis", Proc. Workshop on Independent Component Analysis and Signal Separation (ICA99); Aussois, France, pp. 37-42.

[19] Reza Farrahi Moghaddam and Mohamed Cheriet, 2008 "EFDM: Restoration of Single-sided Low-quality Document Images", Proceedings of ICFHR 2008, Montreal, Quebec, Canada, pp. 204-209.