Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department Of CSE, JayaShriram Group Of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Restoring Degraded Documents by Using Neural Network - KSOM Based Hybrid Techniques

S.Nanthini[1] and M.Yuvarani[2]

Department of ECE, Erode Builder Educational Trust's Group of Institutions, kangeyam, Tamil Nadu, India,[1,2]

Abstract: The objective of this paper is to specify a special problem in character recognition from document images where the verso side scripts appear as noise on front side. 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 in order to improve readability. We propose a new hybrid technique which is based on Neural Network Kohenen Self organizing Map (NNKSOM). The proposed method proves to perform processing techniques with improved performance. This method is absolutely useful for researchers engaged in recognizing any script worldwide as the same kind of distortion can be found in any image used worldwide.

Keywords: Distortion, Bleed-through, Show-through

I.   INTRODUCTION

Ancient documents, Property documents and the like are scanned and converted to digital documents to store them for future use. The scanned images might not be legible due to poor paper quality, spreading and flaking of ink etc.There are many solutions available to restore the characters from these degraded documents. But to be more efficient they need neat and readable inputs. The accuracy of today's document recognition methods fail abruptly when document image quality is poor 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]. In order to improve the performance, the proposed method combines algorithms such as Diffusion Method (DM), Independent Component Analysis (ICA) Double Sided-Flow Based Diffusion Method (DFDM) and NNKSOM. This is one of the most challenging problems in OCR (Optical Character Recognition).
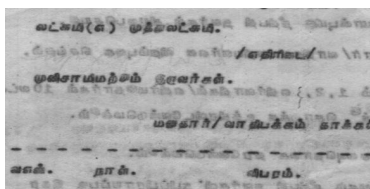

Fig.1. Double-sided distorted document

II.   REVIEW OF LITERATURE

Numerous methods have been proposed to recognize Bleed-though problems. In order to reach the desired goal, an

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department Of CSE, JayaShriram Group Of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

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 Emmanuelle 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.

III.   ALGORITHMS USED

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

3.1 STATISTICAL ALGORITHM

Blind Signal Separation (BSS) application holds a remarkable place in statistical approach. In general BSS problem often referred as 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. 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:
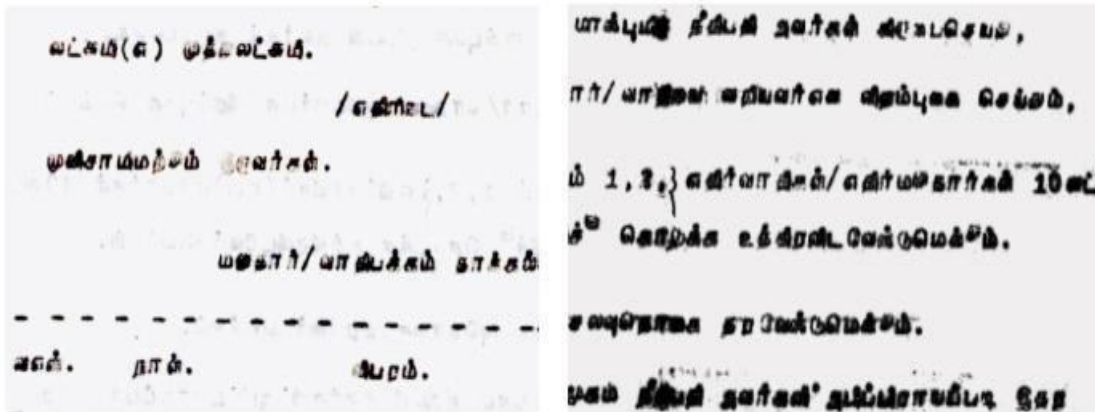
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department Of CSE, JayaShriram Group Of Institutions, Tirupur, Tamilnadu, India on 6ᵗʰ & 7ᵗʰ March 2014

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

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.

3.2.2 Drawbacks 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.

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 neighboring 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 neighboring 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

- DFDM method cancels out the effects of real physical distortion process that occurs over time.
- Additional diffusion processes actually separate the recto and verso side information to the background.
- 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

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.

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.
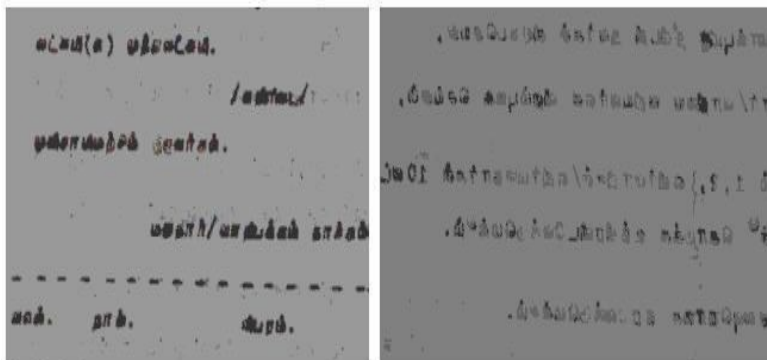
(a) Front Page   (b) Back Page

Fig.3. Resultant Image after Applying Diffusion

IV. 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.

4.1 PROPOSED HYBRID ALGORITHM

Algorithm: Double_Sided_Restoration

Step 1: Implement any diffusion method over input image
Step 2: Implement ICA method over DM_IMAGE
Step 3: Name the resulted image as ICA_IMAGE_1 and ICA_IMAGE_2
Step 4: Implement DFDM method over ICA_IMAGE_1 and ICA_IMAGE_2
Step 5: Use NN technique, to classify information for ICA and DFDM
Step 6: NN makes training for recognition
Step 7: Recognition can be done on the Content-based information
Step 8: Results restore or enhance the input document image

4.2 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.

4.3 ENHANCEMENT

DFDM is a powerful tool for the enhancement and source separation. In general, it requires some source dominance in

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

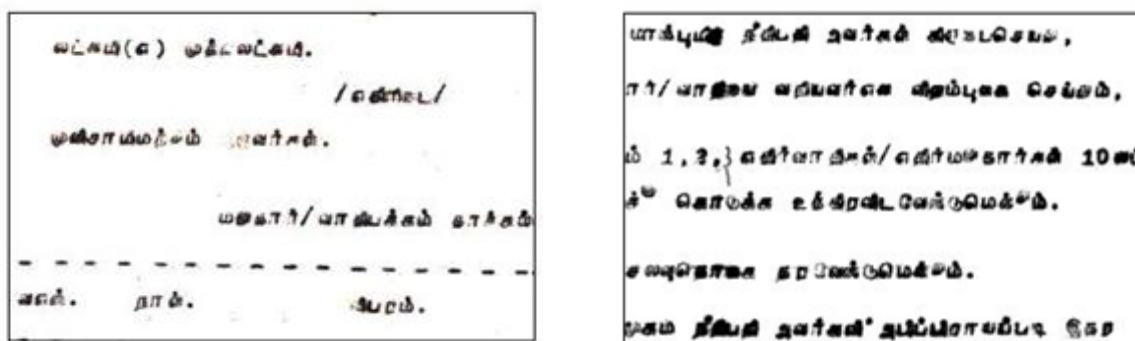**Department Of CSE, JayaShriram Group Of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014**

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.

4.4 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.

$$DM + ICA + DFDM + NNKSOM \qquad (6)$$

The proposed hybrid algorithm is implemented over Fig.1. The resultant output is shown in Fig. 5(a) and (b).

a) Front Page   b) Back Page

Fig.5. Resultant Image by Proposed hybrid technique

V. RESULTS AND INFERENCES

S.NO	TECHNIQUE	MSE	PSNR
1.	Diffusion method	1.9645e+003	15.1984
2.	Ica	1.6672e+003	15.9109
3.	Proposed method	1.1821e+003	17.4041

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department Of CSE, JayaShriram Group Of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

VI. 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
[5] Techniques", ICIP 2007, IEEE conference on Image
[6] Processing, , San Antonio, TX, Vol. 3, pp. III-233-III-236.
[7] Leedham, G., Varma, S., Patankar, A., and Govindaraju, V.,
[8] 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.
[9] Emanuele Salerno, Anna Tonazzini, and Luigi Bedini, 2006,
[10] "Digital image analysis to enhance underwritten text in the archimedes palimpsest", IJDAR, Vol. 9, No. 2-4, pp.79–87.
[11] Xiaowei Zhang, Jianming Lu, and Takashi Yahagi, 2007,
[12] "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
[14] Chee, and Josephine Chang, 2000, "Removal of interfering
[15] strokes in double-sided document images," Applica-tion of Computer Vision, 5th IEEE Workshop Ruini Cao, Ed., pp.16–21.
[16] 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.
[17] Vigario, R.N., 1997, "Extraction of Ocular Artifacts from EEG using Independent Component Analysis", Electroence-phalograph. Clin. Neurophysiol, Vol. 103, pp. 395-404.
[18] Biswal, B.B., and Ulmer, J.L., 1999, "Blind Source
[19] Separation of Multiple Signal Sources of MRI Data Sets
[20] Using Independent Component Analysis," J. Comput. Assist.
[21] Tomogr., Vol. 23, pp. 265-271.
[22] Park, H. M., Jung, H.Y., Lee, T.W., and Lee, S.Y., 1999, "Subband-based Blind Signal Separation for Noisy Speech
[23] Recognition", Electronics Lett., Vol. 35, pp. 2011-2012.
[24] Ypma, A., Pajunen, P., 1999, "Rotating Machine Vibration
[25] Analysis with Second-order Independent Component
[26] Analysis", Proc. Workshop on Independent Component
[27] Analysis and Signal Separation (ICA99); Aussois, France, pp. 37-42.