A Recursive Approach For Bleed-Through Removal

F. DRIRA H. EMPTOZ

LIRIS - INSA de LYON - Bât Jules Verne
20 avenue Albert Einstein 69621 Villeurbanne Cedex France
{fdrira, hubert.emptoz}@liris.cnrs.fr

Abstract

Historical documents are valuable resources worth to be preserved in order to support our cultural and social knowledge. Unfortunately, these supports based on fragile materials are often affected by several types of degradations. Applying restoration techniques on degraded captured digital images of historical documents may be a quick and efficient way to preserve the document and avoid the loss in its content.

This paper presents a new method to restore a particular type of degradation which is referred to as "bleed-through". This degradation is caused by the interference of characters from the reverse side with the text to be read. Our proposed method is based on a recursive approach that relies on two types of analysis: the Principal Component Analysis and the k-means clustering algorithm. The aim here is to extract clear textural images from these interfering and overlapping areas of text. Our restoration method analyses the front side image alone and corrects the unneeded image components. This paper concludes with some experimental results that demonstrate the effectiveness of our proposed method.

1. Introduction

The advance of digital technologies and techniques makes it possible to preserve heritage documents for a longer period. This digital mean of preservation would also enable a widespread diffusion of these documents. Indeed, recent techniques help in producing digital copies of original heritage documents. However, the quality of these digital copies depends greatly on the quality of the original ancient documents. These are often affected by several types of degradations. Baird [1] suggests the following definition of the term degradation (or defect): "every sort of less-than ideal properties of real document images". In fact, old documents, supported by fragile materials, are easily affected by bad environmental conditions. Manipulations, humidity and unfitted storage for many years affect heritage documents and make them difficult to read. Moreover, the digitizing techniques used in image scanning inevitably further degrade the quality of the document images. Indeed, degradations affect ancient documents and make them difficult to read. Resorting to restoration techniques for these deteriorated old documents becomes an increasingly urgent need. Restoration refers to the treatment of a low quality historical document. Restoration techniques can improve the quality of the digital copy of the originally degraded document, thus improving human readability and allowing further application of image processing techniques such as segmentation and character recognition. A large number of algorithms have been developed by the community. However, each of these methods depends on a certain context of use and is intended to process a precise type of defects.

In this study, we will focus on a particular type of degradation, which is referred to as "bleed-through". This degradation is due to ink's seeping through the pages of documents after long periods of storage. The result is that characters from the reverse side appear as noise on the front side. This can deteriorate the legibility of the document if the interference acts in a significant way. An overview of some restoration techniques tackling this kind of degradation is presented in the first section. In the second section, we propose a new algorithm trying to restore such kind of degraded documents by extracting clear textual images from interfering and overlapping areas. The main idea behind our algorithm is to classify each pixel of the page to be processed in one of the following three classes: (1) background, (2) original text, or (3) interfering text. Our problem is therefore close to a segmentation problem. We propose to perform, recursively, a k-means algorithm on the dimensionally reduced image data. This dimension reduction is done through Principal Component Analysis (PCA). Our recursive restoration method does not require specific input devices or the digital processing of the backside to be input. It is able to correct unneeded image components through analysis of the front side image

alone. The third section shows experimental results that verify the effectiveness of our proposed method.

2. Brief review of bleed-through restoration techniques

Thresholding techniques are a simple possibility but remain insufficient for too degraded documents. For instance, Leedham and al. [2] compared several thresholding techniques for separating text and background in degraded, historical documents. The results prove that neither global nor local thresholding techniques perform satisfactorily. Indeed, many restoration approaches dealing with "bleed-through" removal were proposed. Some of them have successfully resolved this problem but under specific conditions. These methods can be divided into two classes according to the presence of the verso side page document: non-blind ones—treating this interference problem using both sides of the document— and blind ones treating this problem without the verso side.

2.1. Non-blind restoration techniques

The main idea of non-blind approaches is mainly based on the comparison between the front and back page, which requires a registration of two sides of the document in order to identify the interfering strokes to be eliminated. Techniques of this type are reported in [3, 4, 5] for scanned documents. Sharma's approach [3] simplifies the physical model of these effects to derive a linear mathematical model and then defines an adaptive linear-filtering scheme. Another approach proposed by Dubois and Pathak [4] is mainly based on processing both sides of a gray-level manuscript simultaneously using a six-parameter affine transformation to register the two sides. Once the two sides have been correctly registered, areas consisting primarily of "bleed-through" are identified using a thresholding technique and replaced by the background color or intensity. In [5], a wavelet reconstruction process is applied to iteratively enhance the foreground strokes and smear the interfering strokes. Doing so strengthens the discriminating capability of an improved Canny edge detector against the interfering strokes.

All these different non-blind restoration techniques dealt successfully with "bleed-through" removal. Nevertheless, a registration process of both sides of the document is required. Perfect registration, however, is difficult to achieve. This is due to (1) different document skews and (2) different resolutions during image capture of both sides, (3) non-availability of the reverse side and (4) warped pages resulting from the scanning of thick documents. The main drawback of this approach is therefore its dependency on both sides of the documents that must been processed together. Resorting to a blind restoration method, i.e. removing the bleed through without the need of the both sides of the document is often a more interesting solution.

2.2. Blind restoration techniques

A variety of techniques have been proposed in this regard. An interesting approach successfully used is based on steered filters. This approach is especially designed for old handwritten documents. A restoration method [6] proposed by Tan and al. consists in adopting an edge detection algorithm together with an orientation filter to extract the foreground edges and remove the reverse side edges. This algorithm performs well and improves greatly the appearance of the original documents. However, one problem with this method is mainly obtained when the interference is so serious that the edges of the interfering strokes are even stronger than that of the foreground edges. As a result, the edges of the interfering strokes would remain in the resultant text image. Another approach proposed by Wang et al. [7] uses directional wavelets to remove images of interfering strokes. The writing style of the document determines the differences between the orientations of the foreground and the interfering strokes. Basically, the foreground and the interfering strokes are slanting along the directions of 45° and 135° respectively. The directional aspect of the transform is capable of distinguishing the foreground and reverse side strokes and effectively removing the appearing interference. This approach produces very interesting results but it remains applicable only to particular cases of character orientation (45° and 135°). All the techniques cited above treat a particular case of degraded document, where foreground and interfering strokes characters are oriented differently, which is not always the case. Other more flexible techniques exist, among which, we can cite techniques based on Independent Component Analysis [8], adaptive binarization [9], self-organizing maps [10], color analysis [11].

So far, we presented a classification of some methodologies proposed to tackle the "bleed-through" degradation. After this short outline, our choice will directed to a blind restoration method as the verso side is not necessary available.

3. The proposed method

3.1. Justification

As already said, the scanned image of a document, which has been subject to "bleed-through" degradation, contains the content of the original side combined with the content of the reverse side. A representative part of such a degraded document (provided by Chatillon-Chalaronne) is shown in Figure 1. Our problem, illustrated in this figure, is to extract clear text strings of the front side from this noisy background. We propose to proceed with a segmentation approach. In fact, the main idea behind our algorithm is to classify the pixels of the page into three classes: (1) background, (2) original text, or (3) interfering text. This last class must be removed from the original page and replaced by the background color (the average of the detected background pixels for example).

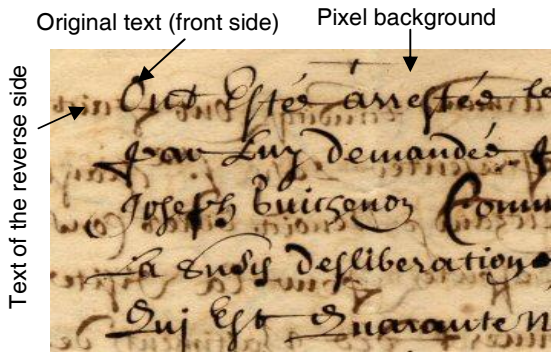

Figure 1: An extract of a degraded document image

"Bleed-through" removal is thus a three-class segmentation problem. Nevertheless, a single clustering step is not sufficient to correctly extract the text of the front side (Figure 2).

Figure 2: Application's Results of the 3-means classification algorithm on a degraded image

Thus, we propose to apply a recursive segmentation method on the reduced data with PCA. To simplify the analysis and reduce its computational complexity, we will restrict ourselves to the case of a two-class problem: original text or not. The proposed method is built then via recursively dividing the test image into two subsets of classes.

The following paragraph will briefly (1) introduce k-means, (2) introduce PCA, and (3) explain the importance of applying k-means on PCA. Our method will be explained in greater detail in the following section (section 4).

(1) k-means is an algorithm [12] using prototypes (centroids) to represent clusters by optimizing the squared error function. The prototypes are initially randomly assigned to a cluster. The k-means clustering proceeds by repeated application of a two-step process where the mean vector for all prototypes in each cluster is computed and then prototypes are reassigned to the cluster whose centre is closest to the prototype. The data points are thus decomposed into disjoint groups such that those belonging to same cluster are similar while others belonging to different clusters are dissimilar.

(2) PCA is an example of eigenvector-based technique which is commonly used for dimensionality reduction and feature extraction of an embedded data. The main justification of dimension reduction is that PCA uses singular value decomposition (SVD) which gives the best low rank approximation to original data. Indeed, PCA can reduce the correlation between the different components where coherent patterns can be detected more clearly.

(3) Applying k-means on PCA: we propose here to apply K-means (K=2) clustering in the Principal Component Analysis (PCA) subspace. Pioneering work [13] has shown that PCA dimension reduction is particularly beneficial for K-means clustering. It was also proven that the continuous solutions of the discrete K-means clustering membership indicators are the data projections on the principal directions (principal eigenvectors of the covariance matrix). More precisely, we decided to apply the segmentation algorithm on image data decorrelated using a PCA. The PCA is computed on the RGB color space. It improves the quality of classification because of its properties which reduce data space and eliminate associations between data. In representing the document image in a convenient vector space, we will succeed to improve the gathering of elements with approximately similar values in order to make them converging to significant classes.

3.2. Description of the method

A new framework based on a recursive approach is presented here, which relies on two types of analysis:

the Principal Component Analysis and the k-means algorithm applied recursively on each generated data image. A scheme of our approach is given in Figure 3. The following steps are performed recursively:

(1) The dimension of an image is reduced and its data is decorrelated using Principal Component Analysis.

(2) The k-means algorithm is applied with parameter k=2, resulting in two classes of image pixels.

(3) The pixels of each class back-projected into the original color space. Each generated class is recursively used as input to the same algorithm beginning with step 1.

The dimension reduction step projects the document image from the original vector space to another reduced subspace generated via PCA. The RGB color space, where each color is represented by a triplet red, green and blue intensity, is used as input.

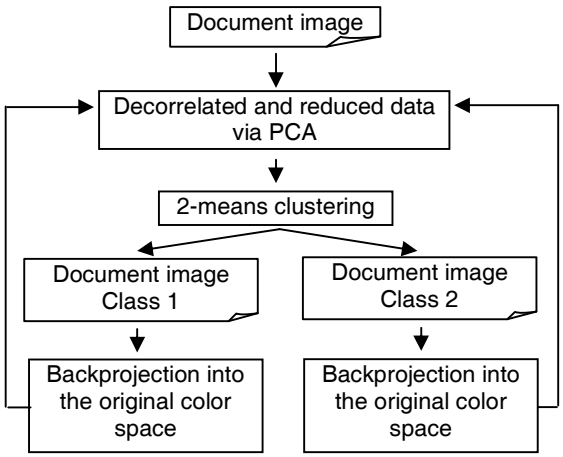

Figure 3: The flowchart of the proposed method

As shown in Figure 4, the first principal component (its eigenvalue represents 99.02% of the total eigenvalues variance) gives a good approximation of the image compared to the other principal components. For instance, when we project onto the directions with biggest variance, we can not only retain as much information as possible but also we can deliberately drop out directions with small variance.

Indeed, selecting the most significant principal components as input to the k-means clustering algorithm reduces the data enough in order to make the problem manageable while at the same time retaining enough information to perform a successful separation. The Figure 5 shows more clearly the difference between the luminance axis and the first component of the PCA. This analysis can better maximize pixel separation projected on its first component.

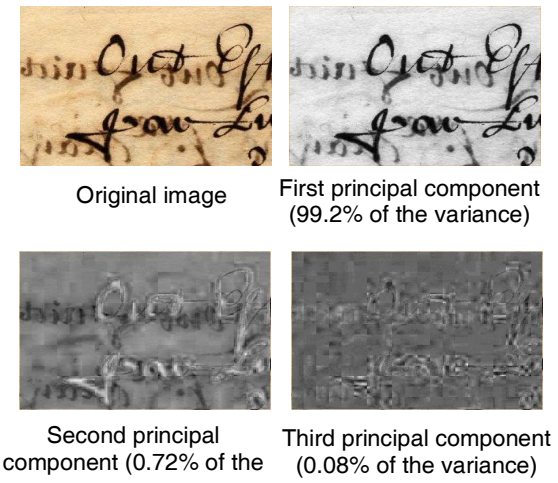

Figure 4: Results of PCA projection

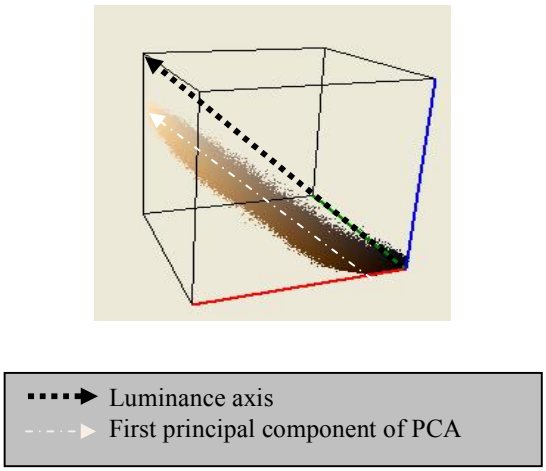

Figure 5: Color distribution of Figure 1 in the RGB cube

This method starts with the whole image set as a single cluster. Then, it is partitioned into disjoint subsets a1 and a2, where the inter-cluster distance is maximized. The subsets a1 and a2 are further subdivided into a11 and a12 and a21 and a22, and so on. The process thus generates a binary tree. One of its leaves represents the expected image which contains the original text. Figure 6 represents an extract of this tree. Only subsets leading to the expected image are represented. As shown in Figure 6, image (a122) is the expected result. The number of iterations in our method has been determined empirically and set to a fixed number of iterations (=3). The result of the algorithm is a set of classes (the leaves of the tree of recursive function calls), where one class represents the pixels of the original handwriting. We can so notice that the

segmentation of the data in a recursive way allows us to refine the final restoration result as soon as we traverse down the binary tree. Our method outperforms other methods that involve a global classification in K classes applied to the entire image (Figure 2). It converges more correctly to the final result.
For the moment, the class containing the original recto side is chosen iteratively by an operator. We are currently working on an automatic detection of this class. While the manual intervention choosing the final result would appear as a weak point in our method, we consider this intervention crucial to preserve the originality of the restored document.

Figure 6: An extract of the generated tree with our proposed method applied on a test image

4. Experimental results and discussion

Experiments were carried out to evaluate the performance of our approach. We used some samples of degraded image documents. An example of a restored image resulting from the application of our method on a degraded document image (Figure 1) is given in Figure 7. This figure shows one of the subsets generated after three iterations of the method. Moreover, this subset represents the front side text and we clearly notice, compared with the test image (Figure 1) that the interfering text has been successfully removed and replaced by the average of the detected background pixels.

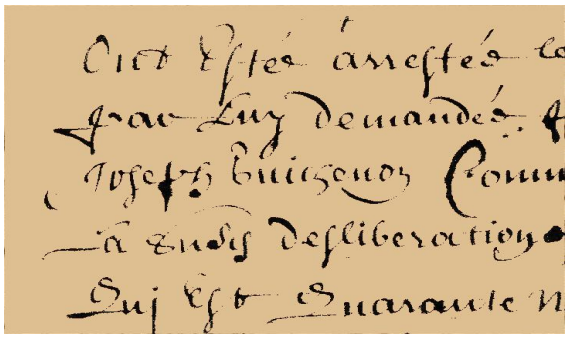

Figure 7: The obtained image with our method applied on figure 1 (after 3 iterations)

We have also validated our method on degraded document images containing black and red text. An example of such document image (provided by the French Institute of Research on Text History (IRHT)) is illustrated in Figure 8. Figures 9 and 10 show the results of the experiment performed on this document image. Those images could be combined to create a restored version (Figure 11) of the degraded image. Other degraded documents (Figure 12, Figure 14 and Figure 16) and their respectively restored versions (Figure 13, Figure 15 and Figure 17) are given below.

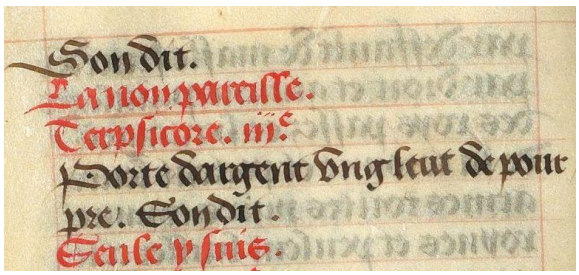

Figure 8: A degraded document image

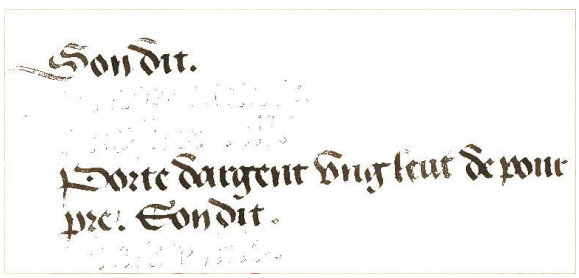

Figure 9: Black text extracted by our method

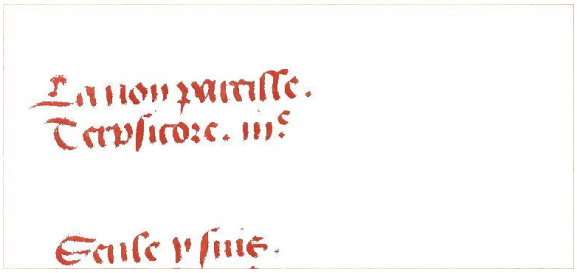

Figure 10: Red text extracted by our method

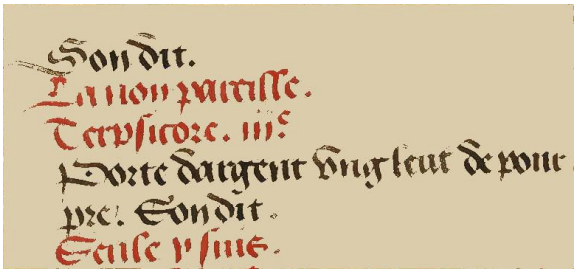

Figure 11: Restored image

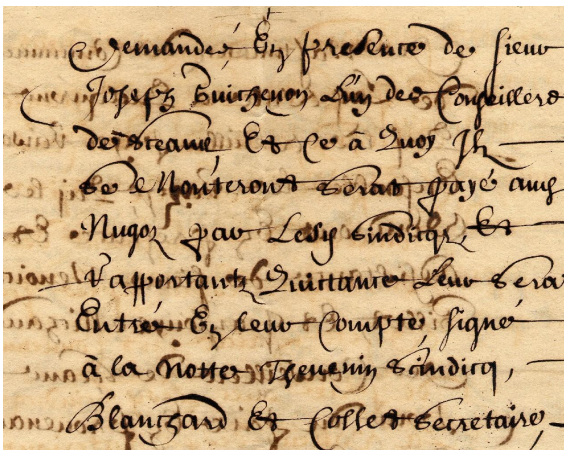

Figure 12: A degraded document image

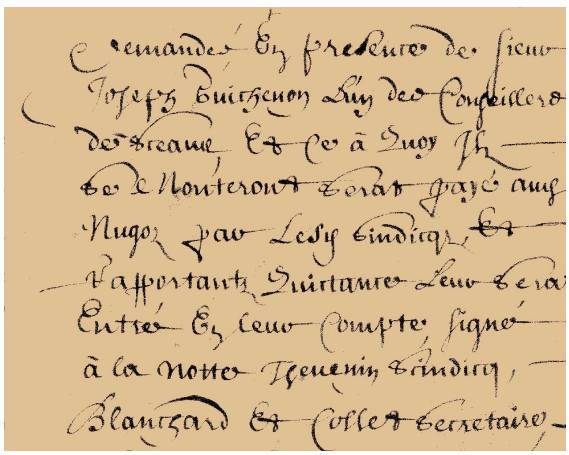

Figure 13: Restored image

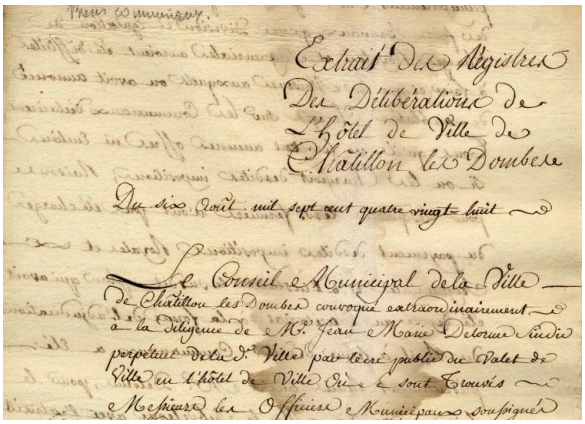

Figure 14: A degraded document image

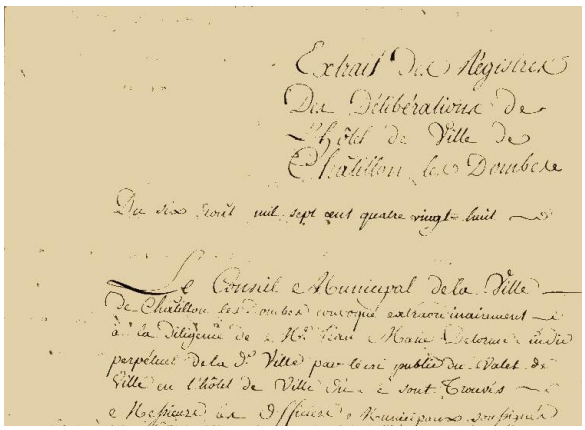

Figure 15: Restored image

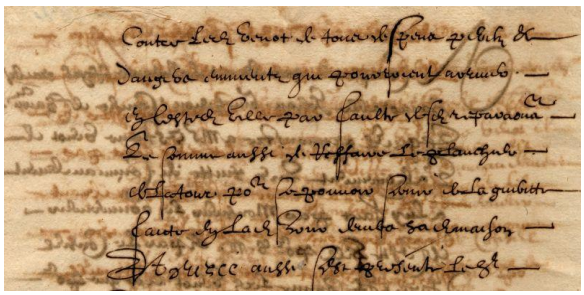

Figure 16: A degraded document image

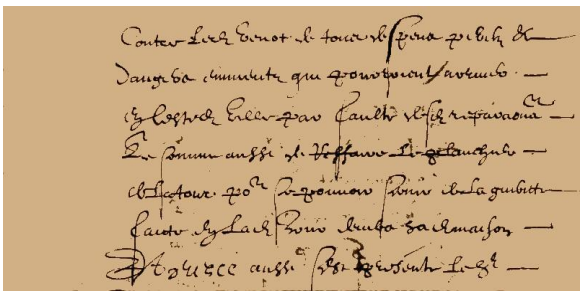

Figure 17: Restored image

Experimental results illustrate the significant performance of this recursive approach, which produces similar results as those obtained with the approach [11] (Figures 18 and 19 are the restored version of an extract of Figure 1). However, it is not hampered by the same restrictions. The approach [11] represents an adaptive segmentation algorithm suited for color document images analysis. It is based on the serialization of the k-means algorithm.

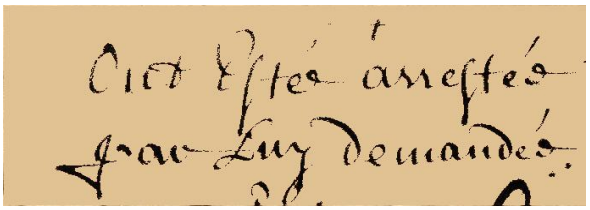

Figure 18: The obtained image with our method

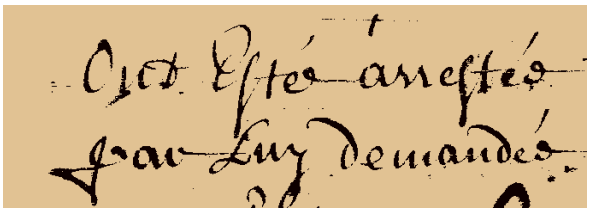

Figure 19: The obtained image with the approach [11]

Compared to other existing methods, our method
(1) does not require specific input devices or the processing of the reverse side of the document to be input. It is able to correct unneeded image components through the analysis of the front side image alone. Our approach can be classified among "blind bleed-through" removal techniques.
(2) does not require any specific learning process such as the case of the self-organizing Maps based approach [10] where a learning process must be performed on each chosen image.
(3) does not require any input parameters as in the case of the serialized k-means based approach. Certainly, this approach gives good results but it is an unsupervised one as the choice of some parameters such as the number of clusters and the color samples for each class are not done automatically.

Nevertheless, one of major weakness of our algorithm resides in its computationally complexity especially for high dimension document images. Indeed, one of our future aims is to obtain a correct front page image with efficient computation and reduced memory demands. This can be done by controlling the recursive decomposition of our original image. In other word, the scope is to allow decomposition only of leaves images that can lead to the corrected expected image. If we suppose that the original text is the darkest one, the analysis of the image histogram values could be a solution. By doing so, we can reach an automatic final class image detection

Moreover, other extensions and refinements of our work will be directed towards the mixture of the obtained image results with further image processing techniques. These techniques could improve the obtained results and help to produce a more readable and visible text. For instance, removal of the interference text may damage the touching characters in the overlapping area. This is due to the fact that some parts of the removed segments possibly belong to both original and interfering text. Broken characters could also been justified by the irregularity of ink color and also the variability of the ink layer's depth over the different characters. This distinguishes handwritten documents from printed ones and makes their treatment more complex. The aim here is to recover broken edges of the words or characters on the front side. Further research will concentrate on these ideas.

5. Conclusion

Both Principal Component Analysis (PCA) and K-means can be combined to make together a powerful tool for image processing tasks. In this paper, they are applied recursively to separate original text from interfering and overlapping areas of text. Experimental results illustrate visual improvement results of digital degraded document images. Certainly, PCA used as a space reduction technique has proven to be powerful as a pre-processing step for the k-means classification algorithm. Nevertheless, the linearity of this transform could limit its application since this transform could not detect at all times the different structures in a given image. Resorting to a suitable nonlinear transform could give better results or decrease the iteration number of the process. Moreover, the choice of the k-means and the PCA, widely used techniques in the literature, represents a first step for testing its relevance. Our future research will investigate other techniques and compare the results with those obtained here to evaluate performances.

6. Bibliography

[1] H. S. Baird, *State of the Art of Document Image Degradation Modelling*, invited talk, IAPR 2000 Workshop on Document Analysis Systems, Brazil, December 2000.

[2] G. Leedham, S. Varma, A. Patankar, V. Govindaraju, *Separating text and background in degraded document images – a comparison of global thresholding techniques for multi-stage thresholding*. In: Proceedings of the 8th international workshop on frontiers in handwriting recognition, pp 244–249, Canada, August 2002,

[3] G. SHARMA, *Cancellation of show-through in duplex scanning*, International Conference on Image Processing (ICIP), vol. 2, pp. 609-612, September 2000.

[4] E. Dubois, A. Pathak, *Reduction of bleed-through in scanned manuscripts documents*, In: Proceedings of the IS&T conference on image processing, image quality, image capture systems, Montreal, Canada, April 2001, pp 177–180

[5] C. L. Tan, R. Cao, P. Shen, *Restoration of Archival Documents Using a Wavelet Technique*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, 1399–1404, October 2002.

[6] C. L. Tan, R. Cao, P Shen, J Chee and J Chang, *Text extraction from historical handwritten documents by edge detection*, 6th International Conference on Control, Automation, Robotics and Vision, ICARCV2000, Singapore, December2000.

[7] Q. Wang, T. Xia, C. L. Tan, L. Li, «Directional Wavelet Approach to Remove Document Image Interference», ICDAR 2003: p736-740, Edinburgh, Scotland, August 2003.

[8] A. Tonazzini, E. Salerno, M. Mochi, L. Bedini, *Bleed-through removal from degraded documents using a color decorrelation method*, DAS 2004, pp 229-240, 2004.

[9] B. Gatos, I. Pratikakis, S. J. Perantonis, *An Adaptive Binarization Technique for Low Quality Historical Documents*, Document Analysis Systems VI, 6th international workshop, DAS2004, pp.102-113, Florence, ITALY, September 2004.

[10] E. Smigiel, A. belaid, H. Hamza, *Self-organizing Maps and Ancient Documents*, Document Analysis Systems VI, 6th international workshop, pp.125-134, Florence, ITALY, September 2004.

[11] Y. Leydier, F. LeBourgeois, H. Emptoz, *Serialized k-means for adaptative color image segmentation – application to document images and others*, DAS 2004, LNCS 3163, pp. 252-263, Florence, Italy, September 2004.

[12] J.A. Hartigan and M.A. Wang. A K-means clustering algorithm. Applied Statistics, 28:100{108, 1979.

[13] D. Chris and H. Xiaofeng. *K-means Clustering via Principal Component Analysis*. Proc. of Int'l Conf. Machine Learning (ICML 2004), Canada. July 2004.