Restoring Ink Bleed-Through Degraded Document Images Using a Recursive Unsupervised Classification Technique

Drira Fadoua, Frank Le Bourgeois, and Hubert Emptoz

LIRIS, INSA de LYON, Bâtiment Jules Verne,
20 Avenue Albert Einstein, 69621 Villeurbanne Cedex, France
{fdrira, Frank.lebourgeois, hubert.emptoz}@liris.cnrs.fr

Abstract. This paper presents a new method to restore a particular type of degradation related to ancient document images. This degradation, referred to as "bleed-through", is due to the paper porosity, the chemical quality of the ink, or the conditions of digitalization. It appears as marks degrading the readability of the document image. Our purpose consists then in removing these marks to improve readability. The proposed method is based on a recursive unsupervised segmentation approach applied on the decorrelated data space by the principal component analysis. It generates a binary tree that only the leaves images satisfying a certain condition on their logarithmic histogram are processed. Some experiments, done on real ancient document images provided by the archives of "Chatillon-Chalaronne" illustrate the effectiveness of the suggested method.

1 Introduction

Historical documents are of great interest to human being. Nowadays, recent techniques help in producing digital copies of these documents to preserve cultural heritage. Nevertheless, the quality of these digital copies depends greatly on the quality of the original documents. These are often affected by several types of degradations limiting their use. 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. A convenient solution to this problem may be the application of restoration methods on these deteriorated document images. Restoration methods can improve the quality of the digital copy of the originally degraded document image, 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 not only due to ink's seeping through the

H. Bunke and A.L. Spitz (Eds.): DAS 2006, LNCS 3872, pp. 38 – 49, 2006.

pages of documents after long periods of storage, but also due to the paper porosity, to the chemical quality of the ink, or to the conditions of digitalization. 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 document images provided by the archives of "Chatillon-Chalaronne". This algorithm performs, recursively, a k-means algorithm on the decorrelated image data with the Principal Component Analysis (PCA) done on the RGB space. It generates a binary tree that only the leaves images satisfying a certain condition on their logarithmic histogram are processed. 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 and to extract clear textual images from interfering areas through analysis of the front side image alone. The method converges towards the restored version of the degraded document image. The third section shows experimental results that verify the effectiveness of our proposed method.

2 Existing "Bleed-Through" Restoration Methods

Degraded document images, which have been subject to "bleed-through" degradation, contain the content of the original side combined with the content of the reverse side. Examples of such degraded document images provided by "Chatillon-Chalaronne" are shown in the Figure 1. Applying restoration methods on these images could be a solution to extract clear text strings of the front side from this noisy background.

Fig. 1. Examples of "bleed-through" degraded document images

Thresholding techniques are a simple solution for restoring such degradation. Nevertheless, these techniques remain insufficient for too degraded document images. For instance, Leedham and al. [2] compared several thresholding techniques for separating text and background in degraded historical document images. The results prove that neither global nor local thresholding techniques perform satisfactorily.

Indeed, looking for efficient restoration methods becomes an urgent need. Some restoration methods 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.

The main idea of non-blind approaches is mainly based on the comparison between the front and back page, which requires a registration of the two sides of the document in order to identify the interfering strokes to be eliminated. Examples of as such approaches are reported in [3, 4, and 5]. 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.

For blind restoration approaches, the restoration process occurs without the verso side. An interesting successfully used approach is based on steered filters. This approach is especially designed for old handwritten document images. A restoration approach [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 approach performs well and improves greatly the appearance of the original document image but it is less or more inefficient when the interference is so serious. In this case of interference, 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 image, 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 Proposed Method

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). "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 (Fig.2). Thus, we propose to apply a recursive segmentation method on the decorrelated data with Principal Component Analysis. 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.

Fig. 2. Results of the 3-means classification algorithm on a degraded image; Top: An extract of a degraded document image; Bottom: Left: Image class n°1, Middle: Image class n°2, Right: Image class n°3

3.1 Justification

The following paragraph will briefly (1) introduce k-means, (2) introduce PCA, (3) explain the importance of applying k-means on PCA, and (4) introduce the logarithmic histogram.

(1) k-means is an algorithm [12] using prototypes 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 or Principal Component Analysis 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) We propose here to apply k-means (K=2) clustering on the Principal Component Analysis subspace. Pioneering work [13] has shown that PCA dimension reduction is particularly beneficial for K-means clustering. 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.

(4) The logarithmic histogram is a histogram with logarithmic scale. We choose this technique as it is common way to scale histograms for display and then assume that a wide range of luminance values can be clearly represented.

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 (PCA) and the k-means algorithm applied recursively on selected 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 PCA.

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

(3) The pixels of each class backprojected into the original color space.

(4) The logarithmic histogram of each class image is printed. A comparison between the two class image histograms is established. The one having more dark pixel values is the one associated to the class image that will be used as input to the same algorithm beginning with step1.

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.

As shown in Figure 4, the first principal component 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

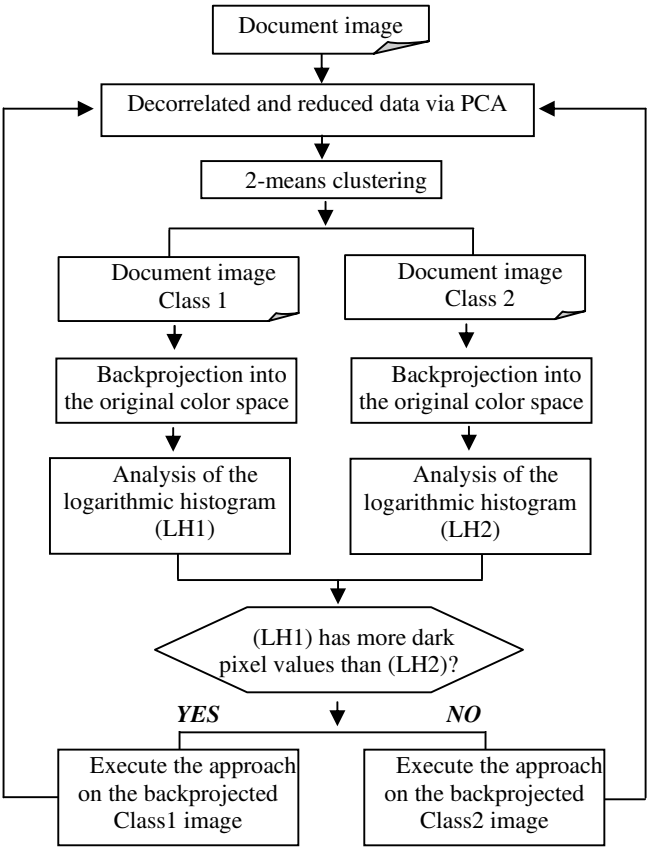

Fig. 3. The flowchart of the proposed method

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.

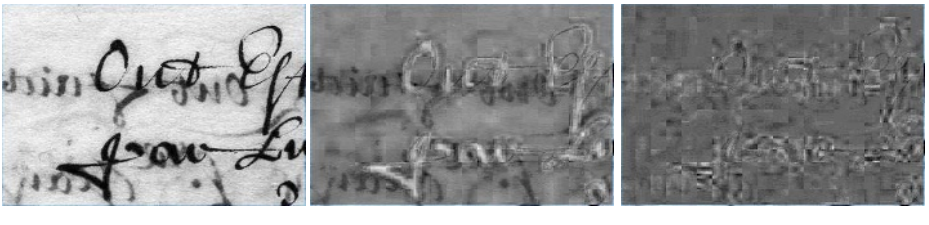

Fig. 4. Results of PCA projection; Left: First principal component (99.2% of the total eigenvalues variance); Middle: Second principal component (0.72% of the total eigenvalues variance); Right: Third principal component (0.08% of the total eigenvalues variance)

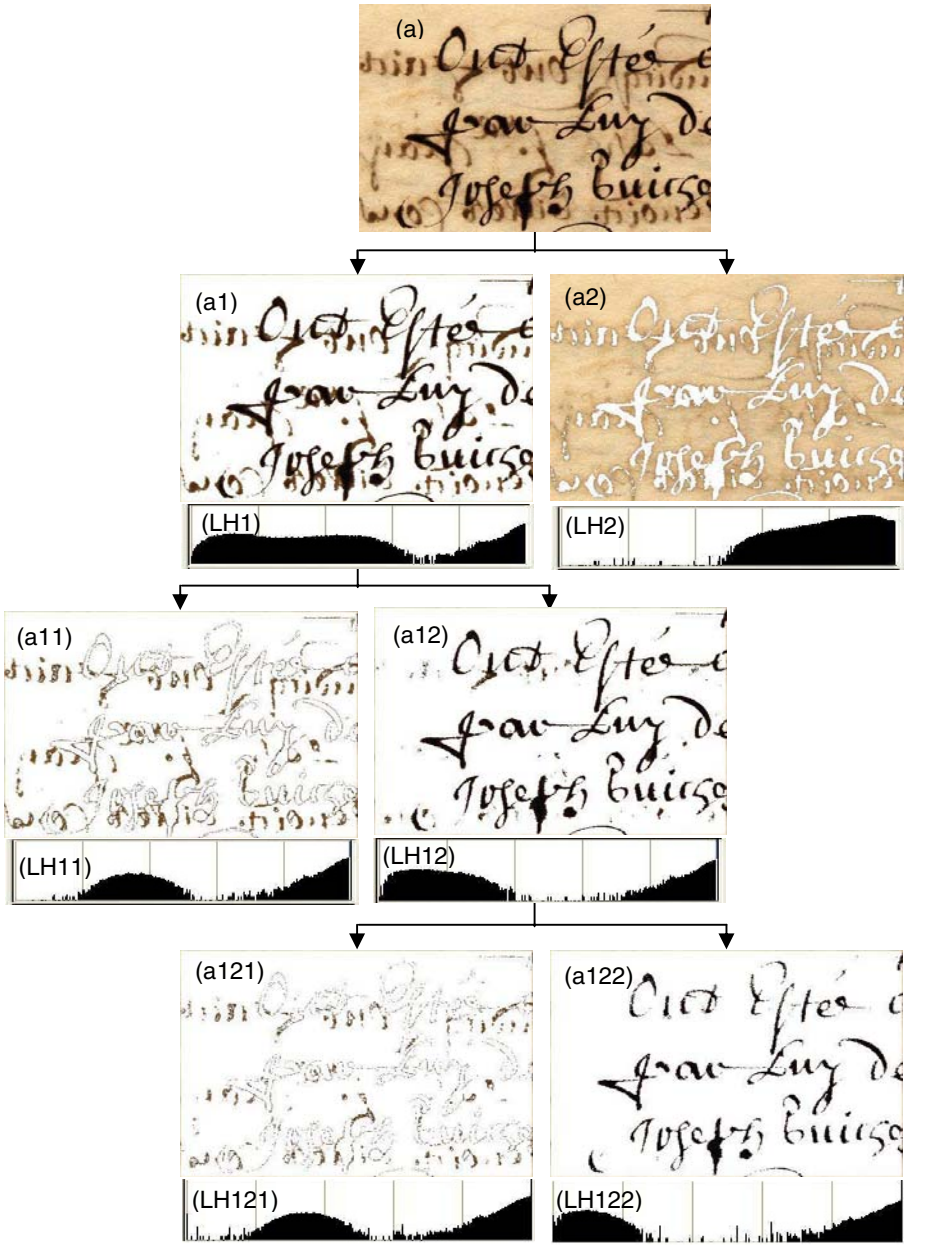

Fig. 5. An extract of the generated tree with our proposed method applied on an extract of a degraded document image (a); (a1), (a2), (a11), (a12), (a121), and (a122) are the different internal nodes of the tree; (LH1), (LH2), (LH11), and (LH12) the logarithmic histogram of (a1), (a2), (a11), and (a12) respectively

The proposed 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 then analyzed by studying their logarithmic histogram. The scope here is to allow decomposition only of the leaves images that can lead to the corrected expected image. If we suppose that the text of the original front side is darker than any other interfering text, the analysis of the image logarithmic histogram values could be a solution. This criterion of selection was followed upon the study of the different provided "Chatillon-Chalaronne" document images. By doing so, we can reach automatic final class image detection. Indeed, the subset image corresponding to the logarithmic histogram having the more dark pixel values is further subdivided and so on for the new generated subsets. The process thus leads after a certain number of iterations to two leaves images that one them represents the expected image. This image contains the original text. Figure 5 represents an extract of this tree. As shown in this figure, 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.

4 Experimental Results

Experiments were carried out on some samples of degraded image documents to evaluate the performance of our approach (Fig.6, Fig.7). The figure 7 illustrates an example of the restored image resulting from the application of our method on a degraded document image. This figure shows one of the subsets generated after three iterations of the method. This subset represents the front side text and we clearly notice, compared with the test image, that the interfering text has been successfully

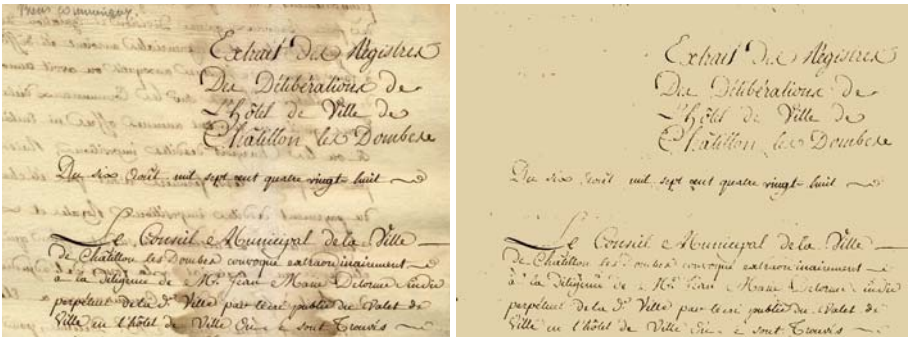

Fig. 6. Left: Scan of an ancient manuscript with "bleed-through" interference provided by Chatillon-Chalaronne. Right: the restored image by the proposed method.

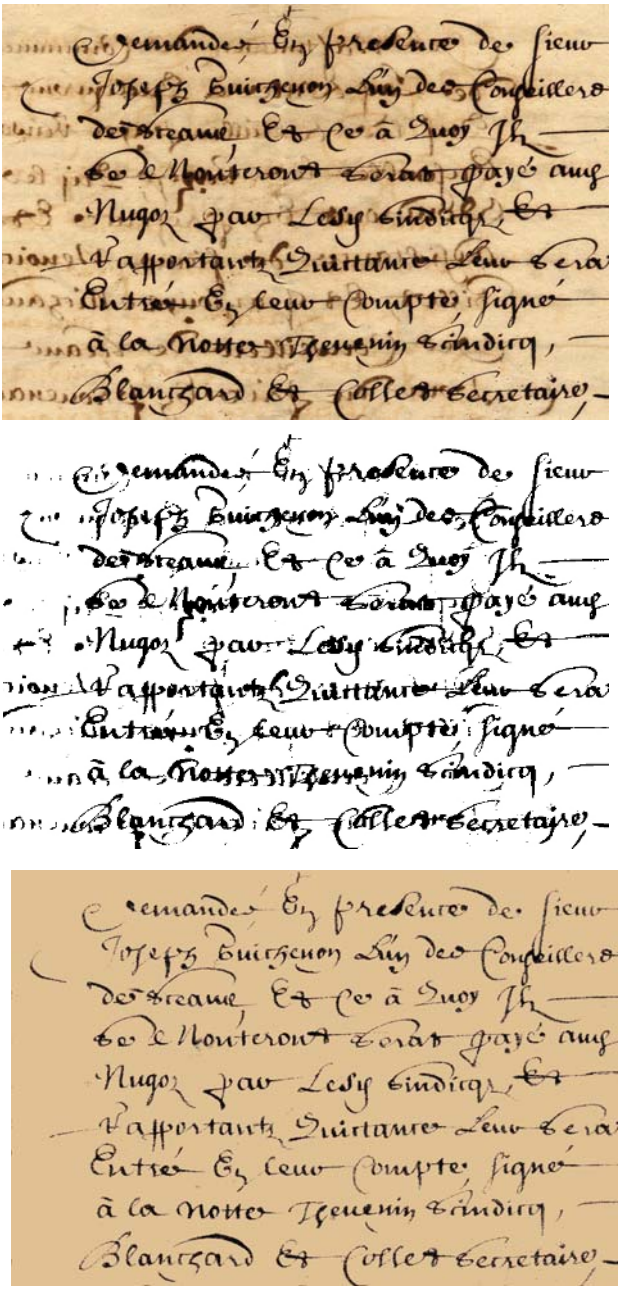

Fig. 7. From top to bottom: Scan of an ancient manuscript with "bleed-through" interference provided by Chatillon-Chalaronne; Application of Sauvola's algorithm; The restored image by the proposed method

removed and replaced by the average of the detected background pixels. This figure illustrates also the fact that classical thresholding techniques such as Sauvola's technique is unable to resolve correctly the problem.

Experimental results illustrate the significant performance of this recursive approach compared to the obtained results of the approach [11] (Fig.8). This approach represents an adaptive segmentation algorithm suited for color document images analysis. It is based on the serialization of the k-means algorithm.

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 approaches.

(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.

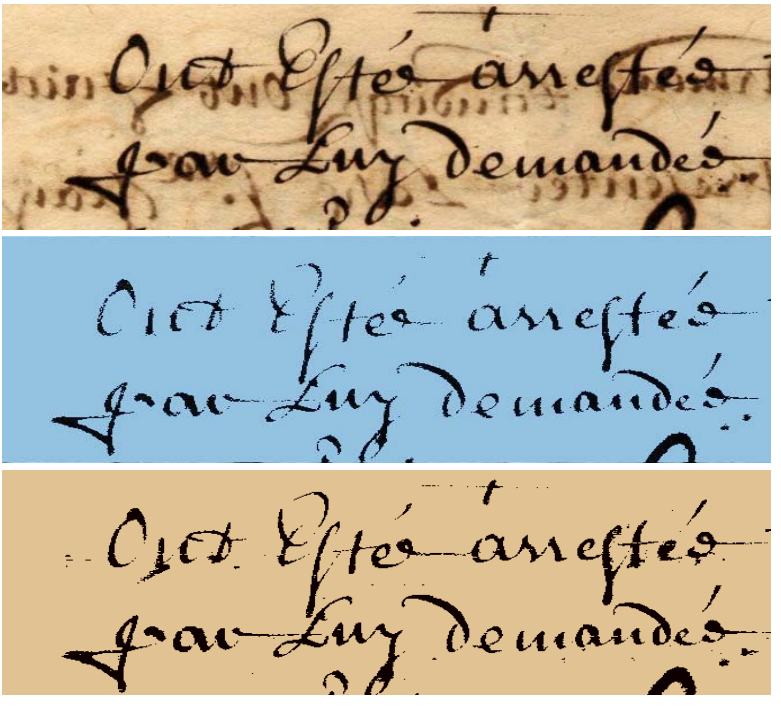

Fig. 8. From top to bottom: Scan of an ancient manuscript with "bleed-through" interference provided by the archives of "Chatillon-Chalaronne"; the obtained image with our proposed method; the obtained image with the approach [11]

5 Conclusions

We demonstrated in this study the effectiveness of our approach in "bleed-through" removal from degraded document images. This approach consists in combining both Principal Component Analysis (PCA) and K-means. These techniques are applied recursively to separate original text from interfering and overlapping areas of text. The stopping criterion for the proposed recursive approach has been determined empirically and set to a fixed number of iterations. Further research will investigate an automatic determination of this criterion.

The analysis of the image logarithmic histogram values is introduced in order to optimize our recursive approach. Thus, we succeed in automatically detecting the final class image representing the restored version of the degraded document image. Successful experiments were done on real ancient document images provided by the archives of "Chatillon-Chalaronne". Other experiments on other archive document images are in progress.

The application of PCA used as a space reduction and data decorrelation technique has proven to be powerful as a pre-processing step for the k-means classification algorithm, however, the linearity of this transform could limit its application. Indeed, 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. 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.

References

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, December 2000.
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.