# Old Manuscripts Restoration Using Segmentation and Texture Inpainting

Insaf Setitra and Abdelkrim Meziane
Research Center on Scientific and Technical Information Cerist, Algiers, Algeria
E-mail: isetitra@cerist.dz, ameziane@mail.cerist.dz

***Abstract*—While manifold works are done in the area of natural image processing to enhance image content, images of documents and more specifically old manuscripts images still suffer from degradations. State of the art approaches show performance in restoring degraded manuscripts images but still a lake is observed in this area. Old manuscripts images are important in that they present a part of a culture for many countries. In this paper we consider restoration of Algerian old manuscripts which present much degradation due to physical conditions and human handling. Restoration is performed by removing spurious details and inpainting missed content. We overview important works done in this area and propose a system to restore content. The system is divided in two modules; the former is concerned with elimination of ink spots, stamps and border notes. The second module is concerned with texture inpainting which repaints missed content. The system provides an interactive user interface, enabling users to choose type of restoration and its parameters. Tests are done on synthetic images and images of real manuscripts.**

***Index Terms*—old manuscript, manuscript restoration, image processing, segmentation, texture inpainting**

## I. Introduction

Although natural image restoration is widely treated in the literature, document image restoration is given less interest. Even though degradation in document and especially old manuscripts are very widespread and need to be recovered by restoration methods [1] and [2].

Among methods of restoration of old manuscripts, we account separation of background from foreground. Such methods include separation by orientation analysis [3], by form analysis [4], by color analysis [5] and [6], by model [7], by prototype [8] and by 2D and 3D models. These methods are stated as supervised methods. Unsupervised methods in another hand include separation of background from foreground by: a separation of source [9], using mathematical morphology [10] and using morphological, colorimetric and texture properties [11].

More recent methods include treating manuscript image in a multichannel space. In [12] the image is multispectral and text extraction is performed in more than three dimensions. Classification of the text and non text is then used to remove spurious content. In [13] authors propose an adaptive statistical method for binarization of historical documents. Binarization is important in that it serves as a separation between important content and less important one.

By removing unimportant content in the manuscript, holes may appear in the image. To deal with this issue, texture inpainting is used to retrieve lost content [14]. In image processing, texture inpainting is defined as the process of applying sophisticated algorithms to reconstruct missed parts of an image using information contained in the border of the degraded zone. [15].

Many approaches are used to reconstruct content by inpainting. This can be done using variation models [14], similarity search models [14] and [15] or by learning based models [7], [16] and [17].

Exemplar based reconstruction method [18] as a focus of one module presented in this paper, is used to complete degraded zones by computing a priority on each patch used for reconstruction. This priority is computed using a data term. The latter is the product of neighboring pixels of the degraded image and the gradient of the image. This method is easy to implement, can reconstruct the image even with sophisticated texture and is fast to compute. In another side, it can lose information when the part to be reconstructed is too big. A complete definition of all previous methods can be found in [14].

In this work, we are interested in Algerian old manuscripts located on the National Center of Manuscrits in Adrar, Algeria. As a wealth of the country, these manuscripts have to be protected and restored. Our aim is to restore virtually these manuscripts by image processing techniques. We process with easy and fast algorithms which we divide into two modules; the first one is used to eliminate spurious content and the second one to inpaint content. For our purpose, segmentation of image manuscript is first done to segment text and non text content. Then non text content is removed and inpainting by exemplar based is used to retrieve lost content. The whole is grouped in our system which provides an interactive interface for user.

The remaining of this paper is organized as follows. Each module is described in Section II and experimental results are presented in Section III. We conclude in Section IV with some critics and perspectives for future work.

Manuscript received July 31, 2013; revised October 23, 2013.


doi: 10.12720/ijoee.2.2.107-111

## II. Our Proposed System

The system of restoration we propose is divided into two primary modules. The first module is used to eliminate superfluous content. The second module uses output of the first one, where details lost during the first stage or present in the original image are recovered.

### A. Superfluous Content Suppression Module

In our case, superfluous content is considered with annotations often stated in the border of the manuscript, ink spots and stamps which may be rectangular or circular. This content is not necessarily superfluous for many applications, but in our case, we considered that the most important content of the manuscript is the one written by its original authors. Stamps for example are added to original documents some centuries after the original writing of the manuscript. Annotations in another hand may be added by someone who has read the manuscript at any time. Fig. 1 shows the flowchart of the module.

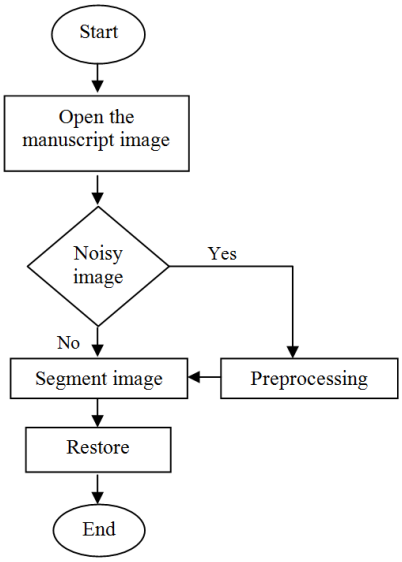


Figure 1. Flowchart of the superfluous content suppression module.

- Preprocessing

This step is performed if the user would like to enhance image quality before performing the restoration. Two enhancement methods are proposed: luminosity enhancement and thresholding.

*Luminosity enhancement:* The luminosity of the original image is enhanced the following way:

Considering the image in a three channel space, each pixel is represented by three components R (red component), G (green component) and B (blue component). For each pixel of the original image, and for each channel of the pixel, the pixel of the enhanced image is calculated as follows:

$$C`_n = C_n + (S * rate)/256 \quad (1)$$

where: C`n is $n^{th}$ component of the enhanced image and n its channels: R, G and B. $C_n$: is the $n^{th}$ component of the original image and n its channels: R, G or B. S is the mean of the three components of the original image: $S=1/n\ \Sigma\ (C_n)$ and rate = (TX*256)/100 (TX is a luminosity parameter).

*Thresholding:*

Thresholding is done following the original Otsu algorithm. Since this method is applied to grey level images and uses histogram of images, we first convert the image to grey level and then compute its histogram.

After this we follow the Otsu algorithm.

- Segmentation with k nearest neighbors

We propose in this paper a very simple algorithm inspired of the classical classification algorithm K Nearest Neighbors KNN. One disadvantage of the classical KNN is the learning phase. We avoid this step by segmenting the image based on a neighboring distance. In our proposed method we define a threshold of neighboring distance between gray level pixels of the image. If the distance between two pixels is less than this threshold then both pixels belong to the same class, otherwise one of them belongs to another class. The only parameter here is the neighboring distance which we chose empirically.

We support our method by comparing it to K means clustering. A disadvantage of Kmeans is that the number of classes has to be defined. Classes in our case are: essential content and spurious content. Spurious content may be annotations, stamps, ink spots and other types of degradation. Some manuscripts can contain both classes while others may not. We can't decide then which number of classes to choose before launching the Kmeans clustering. We implemented the Kmeans algorithm and compared the results with our approach. Visually, our approach gave better results.

- Classification of segments

In the classification step, we attribute a label to each segment provided by the previous step. Again, classes in our case are: essential content (the original one) and spurious content (stamps, annotations and ink spots).

The rational of classification is that significant content is very often larger than the non-significant one and may appear more often. On this basis, for each segment, we compute the number of pixels which compose it which we call the weight of the segment. The segment with the largest weight is considered essential content and the remaining segments are classified into spurious content.

This method has the advantage of being simple to implement and having a fast execution time.

- Restauration

In the restoration step, all pixels belonging to the spurious content class are eliminated. The image is then divided into background and foreground. Foreground is the essential content provided with the previous step, background is the rest. The mean of the background is then calculated and all pixels except the essential content are replaced with that mean. In the case of a printable version, background pixels are given the white value.

### B. Missed Content Inpainting Module

Content of manuscripts may be lost during many stages and degradation may be physical and logical. To retrieve the lost content due to degradations, many

approaches are proposed. We have used in our work the exemplar based approach to reconstruct our missing content in the manuscript image. Fig. 2 shows the flowchart of this module.

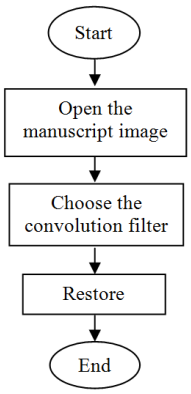


Figure 2. Flowchart of the missed content inpainting module.

Restoration is performed in several steps. These steps are: Select the missing area, locate border pixels and inpaint by exemplar based.

- Select the missing area

Selection of the missing area can be done in two ways: automatically and manually.

In an automatic way, a zone is considered as degraded if its color is given a certain value. This value is chosen according to the type of degradation. Degradations are usually presented as spots and holes. Their color is usually very darker or lighter than the rest of the image and its amount is usually less than other colors. We have chosen in our experiments a blue color as a synthetized degradation and use Euclidian distance to decide whether the pixel is considered as a degradation or not.

- Locate border pixels

A convolution is applied on the missing area to get their border pixels. Convolution can be made using a Laplacian or a Sobel filter.

- Inpaint by exemplar based

Once the border pixels depicted, we can proceed with exemplar based inpainting to reconstruct the missing zone. To do that, first a size of a patch is defined manually, then several steps have to be followed.

Initialize the confidence.

$$\text{Confidence}[x,y]=\begin{cases} 0 \text{ if the pixel } (x,y) \in M \\ \text{Otherwise} \end{cases} \quad (2)$$

where $M$ is the missing area of the image.

Initialize a variable called a data term (we did in our case to 0,1).

Compute the confidence for each pixel in the border of the region to be reconstructed with the formula:

$$Confidence[x,y] = \frac{confidence[x,y]}{N \times M}. \quad (3)$$

where N and M are respectively the height and the width of the chosen patch (a patch is just a rectangle for which we give manually the values of high and width).

Compute the data term with the formula:

$$\begin{aligned} Data[x][y] = {} & GradientX[x][y] \\ & * NX[x][y] \\ & + GradientY[x][y] \\ & * NY[x][y]). \end{aligned} \quad (4)$$

where NX and NY is the gradient of the patch region, and GradientX , GradientY : are the gradients of the original image in the directions X and Y respectively.

Compute the confidence term for each patch with the formula:

$$C(p) = 1 - \omega * confidence[x,y] + \omega. \quad (5)$$

where ω is a parameter assumed to be 0,7 in our case.

Compute a priority for each patch with the formula:

$$P(p) = C(p) + D(p). \quad (6)$$

where C(p) and D(p) are respectively the confidence and the data term of the patch p.

Choose the patch with the higher priority computed previously. It is assumed in this method that the patch with higher priority has the more similar form as the missing area.

Fill the missing area by propagating isophotes corresponding to the patch of higher priority and update confidences.

### C. Experimental Results

Both modules were developed with java language under Eclipse 3.6 platform. Fig. 3 and Fig. 4 represent interfaces of the first and the second module respectively.

The first interface gives the user the ability to choose enhancing the image before restoring it or restoring it directly. And the second interface gives him the ability to select his filter and to use standard inpainting or fast inpainting. Difference between both is the patch where in fast inpaining only neighboring pixels are considered in the first step. In addition, in the second module, user can select the degraded region of the image manuscript he would like to inpaint. Some results of restoration of both methods are presented in figures from 5 to 9.

To compare segmentation using K nearest neighbors, we implemented the K means segmentation, and as shown in Fig. 6(b) and Fig. 7(a), segmentation using K means gave poor results. This is mostly due to the fact that spurious content can have the same properties as the original content, which may false segmentation results with K means. Also, giving 3 means to the algorithm is a bad choice because in this case, only two classes are present in the image; text and stamp. Giving three means conduct to having 3 classes (colored in yellow, black and orange in the image). A Kmeans with three means would give a better result in the case of this image, but would give poor results in an image containing only text and so on. Visually, segmentation with KNN gave better results as the text is well separated from the stamps, however, in the case the stamp is near the text, this method would give poor results. We would investigate this issue in future works.

For retrieving lost content, our experiments showed that a patch of 12x12 gave better results compared to a patch of 4x4 and 6x6. However, we assume that this size is still application and image dependent. Besides, as seen in Fig. 8(b) and Fig. 9, the inpainted content looks better restored while using the Laplacian filter compared to the Sobel filter. However, as the inpainting focus on neighboring pixels instead of using semantic information, the inpainting method can give erroneous results.


Figure 3. Main user interface of the missed content inpainting module.


Figure 4. Main user interface of the missed content inpainting module.

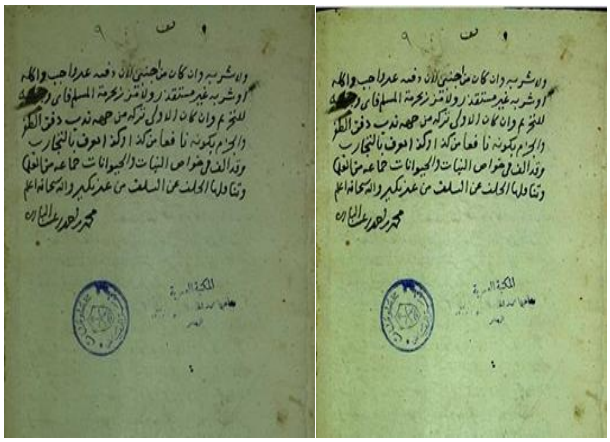

Figure 5. (a) Original manuscript image, (b) Luminosity enhancement of the original manuscript image.

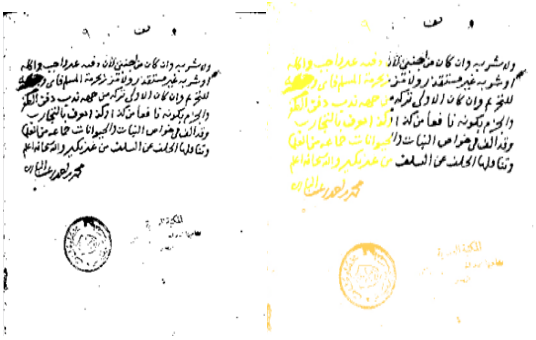

Figure 6. (a) Thresholding of the original manuscript image, (b) Segmentation of the original manuscript image using K means, K=2.

Figure 7. (a) Segmented image using K Nearest neighbors, k=3, (b) output of the first module using background mean.

Figure 8. (a) Output of the first module in a printable version, (b) output of the second module using Laplacian filter.

Figure 9. Output of the second module using a Sobel filter.

## III. CONCLUSION

In this paper we have presented our system for restoration of Algerian manuscripts. We used image preprocessing to enhance luminosity and threshold

images. We then used a K nearest neighbors' segmentation to separate spurious content from original one. And then eliminate this spurious content by a weighting method. Although this method is quite basic, it showed satisfying results in our case. A second step was to inpaint lost content using an exemplar-based technique to find the most similar pixels of the degraded region and then paint the missing texts by following continuity of contours.

As perspectives, we would like to extend our algorithms so that they can cover more manuscripts images as the latters are usually different. We aim also to improve the segmentation method as it does not cover the case where neighboring pixels are themselves spurious content. We would like also to combine some learning based methods so that the inpainting technique can use semantic knowledge to better reconstruct text and have a more coherent content of the manuscripts.

## ACKNOWLEDGMENT

The authors wish to thank Zinnedine Lachib, Wail Cheikh and Cherifa Mokhbat for their valuable effort in accomplishment of this project.

## REFERENCES

[1] R. F. Moghaddam and M. Cheriet, "RSLDI: Restoration of singlesided low-quality document images," *Pattern Recognition*, vol. 42, no. 12, pp. 3355-3364, December 2009.

[2] D. Rivest-Hénault, R. F. Moghaddam, and M. Cheriet, "A local linear level set method for the binarization of degraded historical document images," *International Journal on Document Analysis and Recognition*, vol. 15, pp. 101-124, April 2011.

[3] E. Dubois and A. Pathak, "Reduction of bleed-through in scanned manuscripts documents," in *Proc. IS&T Conference on Image Processing, Image Quality, Image Capture Systems*, vol. 5, April 2001, pp. 177–180.

[4] Q. Wang, T. Xia, C. L. Tan, and L. Li, "Directional wavelet approach to remove document image interference," in *Proc. ICDAR*, Edinburgh, 2003, pp. 736-740.

[5] E. Smigiel, A. Belaid, and H. Hamza, "Self-organizing maps and ancient documents," in *Proc. 6th International Workshop on Document Analysis Systems*, 2004, pp. 125-134.

[6] U. Garain, T. Paquet, and L. Heutte, "On foreground-background separation in low quality document images," *International Journal on Document Analysis and Recognition*, vol. 8, no. 1, pp. 47-63, 2006.

[7] H. S. Baird, "Document image defect models," in *Structured Document Image Analysis*, New York: Springer, 1992, pp. 546-556.

[8] J. D. Hobby and T. K. Ho, "Enhancing degraded document images via bitmap clustering and averaging," in *Proc. 4th International Conference on Document Analysis and Recognition*, 1997, pp. 394-400.

[9] A. Tonazzini, L. Bedini, and E. Salerno, "Independent component analysis for document restoration," *International Journal on Document Analysis and Recognition*, vol. 7, no. 1, pp. 17-27, March 2004.

[10] S. Lefevre and J. Weber, "Automatic building extraction in VHR images using advanced morphological operators," presented at Urban Remote Sensing Joint Event, Paris, France, 11-13 April 2007.

[11] E Aptoula and S. Lefèvre, "Morphological description of color images for content-based image retrieval," *IEEE Transactions on Image Processing*, vol. 18, no. 11, pp. 2505-2517, 2009.

[12] R. Hedjam and M. Cheriet, "Historical document image restoration using multispectral imaging system," *Pattern Recognition*, vol. 46, no. 8, pp. 2297-2312, 2013.

[13] R. Hedjam, R. Moghaddam, and M. Cheriet, "A spatially adaptive statistical method for the binarization of historical manuscripts and degraded document images," *Pattern Recognition*, vol. 44, no. 9, pp. 2184-2196, 2011.

[14] T. Chan and J. Shen, "Mathematical models for local non texture inpaintings," *Siam Journal on Applied Mathematics-SIAMAM*, vol. 62, no. 3, pp. 1019-1043, 2002.

[15] C. Ballester, M. Bertalmio, V. Caselles, and J. Verdera, "Filling-in by joint interpolation of vector fields and gray levels," *IEEE Transactions on Image Processing*, vol. 10, no. 8, pp. 1200–1211, 2001.

[16] A. Efros and T. Leung, "Texture synthesis by non-parametric sampling," in *Proc. International Conference on Computer Vision*, vol. 2, 1999, pp. 1033–1038.

[17] Q. Zheng and T. Kanungo, "Morphological degradation models and their use in document image restoration," in *Proc. International Conference on Image Processing*, 2001, pp. 193-196.

[18] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Trans. on Image Processing*, vol. 13, no. 9, pp. 1200-1212, 2004.

**Insaf Setitra** received the bachelor degree in Information systems and the master degree in Software engineering from the University of Sciences and technology Houari Boumediene Algeria in 2008 and 2010 respectively. She also received a bachelor degree in management from the university of management and economics of Algiers in 2006. She's currently working as a research engineer in the Information and Multimedia systems' Laboratory (DSISM) of the Research Center on Scientific and Technical Information (Cerist) in Algeria. Her current research interests concern image processing, image enhancement and computer vision.

**Abdelkrim Meziane** obtained respectively the Engineer degree in computer science from the University of Science and Technology Houari Bomediene of Algiers in 1986, the DEA from INSA of Lyon (France) in 1987, the Master degree from Spatial Techniques National Center, Arzew, Algeria, in 1996, and the PhD from Oran University of Science and Technology in 2006. He worked as a research assistant at the Spatial Techniques National Center, Arzew, then as an assistant professor at Oran University of Science and Technology, and as researcher at CERIST, a computer center of research in Algiers. He is now the head of the DSISM laboratory at CERIST center. His principal interest research areas include image analysis, hospital information systems, and medical imaging.

 