

# A Concept for the Separation of Foreground/Background in Arabic Historical Manuscripts using Hybrid Methods

W. Boussellaa<sup>1</sup> and H. El Abed<sup>2</sup> and A. Zahour<sup>3</sup>

<sup>1</sup>REsearch Group on Intelligents Machines (REGIM), ENIS, University of Sfax, Tunisia

<sup>3</sup>Institut for Communications Technology (IfN), Braunschweig Technical University, Germany

<sup>2</sup>Equipe Gestion Électronique de Document (GED), University of Le Havre, France

---

## Abstract

*This paper presents a new color document image segmentation system suitable for historical Arabic manuscripts. Our system is composed of a hybrid method which couple together background light intensity normalization algorithm and k-means clustering with maximum likelihood (ML) estimation, for foreground/ background separation. Firstly, the background normalization algorithm performs separation between foreground and background. This foreground is used in later steps. Secondly, our algorithm proceeds on luminance and distort the contrast. These distortions are corrected with a gamma correction and contrast adjustment. Finally, the new enhanced foreground image is segmented to foreground/background on the basis of ML estimation. The initial parameters for the ML method are estimated by k-means clustering algorithm. The segmented image is used to produce a final restored document image.*

*The techniques are tested on a set of Arabic historical manuscripts documents from the National Tunisian Library. The performance of the algorithm is demonstrated on by real color manuscripts distorted with show-through effects, uneven background color and localized spot.*

Categories and Subject Descriptors (according to ACM CCS): I.7.5 [Document and Text Processing]: Document Capture/ Document analysis

---

## 1. Introduction

The historical documents, preserved at the National Library of Tunisia, are considered as an important part of Arabic cultural heritage. These funds suffer from a progressive degradation and therefore risk to disappear. The automatic processing of this type of documents in order to restore and use, is a definite advantage which is confronted with many difficulties due to the storage condition and the complexity of their content. In fact, historical documents have many particularities which hinder classical color document image segmentation algorithms. Figure 1 illustrates the most common deteriorations that appeared in historical Arabic document images which are: The show-through effects (Figure 1 left), the presence of spot due to the humidity absorbed by paper, and an uneven background color paper (Figure 1 right), the presence of fold and tear, and the distortions due to the natural curvature of pages.

Most previous document image enhancement algorithms

have been designed primarily for binarization of modern documents. These methods aim to extract text from noisy documents with uneven background. Three popular methods, namely Otsu's thresholding technique [Ots79], entropy techniques proposed by Kapur and al. [KSW85] and the minimal error technique by Kittler and Illingworth [KI86], are analysed and compared in [LYT\*03, LVPG02]. Another entropy-based method specially designed for historical document segmentation [CR00] deals with the noise inherent in the paper especially in documents written on both sides. Wang and al. [WXL03] presented methods to separate text from background noise and bleed-through text (from the backside of the paper) using direct image matching and directional wavelets. Other methods for historical document image enhancement are driven by the goal of improving human readability of the documents [CR02].

This paper presents a new method for foreground/background segmentation of color historical

Arabic manuscripts. This method combines two techniques of segmentation: The foreground/background separation with background light intensity normalization algorithm and the improvement of the obtained result on the basis of ML estimation after contrast adjustment with gamma correction and histogram normalization of the foreground image. The following paper describes the proposed method and experimental results.



**Figure 1:** Arabic historical documents image: (left) Show-through effects, (right) uneven background

## 2. State of the Art

The works described in [BHH\*98, GPH05, LBE04] are dedicated for foreground/background separation of color document images. DjVu [BHH\*98] implements an efficient foreground/background separation in the context of compression. The approach is based on a multi-scale bi-color clustering that considers several grids of increasing resolution. The technique works well for a large class of documents (gray as well as color) but fails for documents with low contrast.

Leydier and al [LBE04] have achieved an adaptative algorithm for the segmentation of color images suited for document image analysis. The algorithm is based on a serialization of the k-means algorithm that is applied sequentially by using a sliding window over the image. The algorithm reuses information about the clusters computed by the previous classification and automatically adjusts the clusters during the windows displacement in order to better adapt the classifier to any new local modification of the colors. The used colorspaces are RGB and HSL (Hue, Saturation, and Luminosity).

Garain and al [GPH05] have proposed an adaptative method for foreground/background separation in low quality color document images. A connected component labelling is initially implemented to capture the spatially connected similar color pixels. Next, Dominant background components are determined to divide the entire image into a number of grids each representing local uniformity in illumination background. Finally, foreground parts are located using local information around them. This method achieved good results compared to DjVu [BHH\*98].

Shi and al [SG04, SG05] have proposed a color document image enhancement algorithm of palm leaf manuscripts.

This method is based on background light intensity normalization. The background approximation is designed to overcome the unevenness of the document background and the low contrast. The techniques are tested on a set of palm leaf images from various sources and the results show significant improvement in readability.

## 3. Proposed Methodology

The proposed document enhancement methodology permits the improvement of the quality of historical Arabic manuscripts which presented uneven background and low contrast due to the traditional mode of manufacture and the effect of ageing and degradation. It consists of the following steps: foreground extraction, contrast adjustment, foreground/background segmentation, reconstruction of document image with smoothing. The developed document segmentation method operates with background light intensity normalization algorithm proposed by Shi and al [SG04, SG05] and applied to palm leaf manuscripts. We have improved this technique with the histogram normalization used in color image manuscript context. The segmentation method proceeds on luminance and distort the contrast. These distortions are corrected with a gamma correction and contrast adjustment. The new enhanced foreground image is segmented to foreground/background on the basis of ML estimation. The initial parameters for the ML method are estimated by k-means clustering algorithm. The segmented image is used to produce a final restored document image. Figure 2 presents the flowchart of our proposed methodology.

The steps below are described in the following sections:

- Application of an iterative background light intensity normalization algorithm for a first foreground/background separation.
- Correction of visual distortions of obtained foreground using gamma correction and histogram normalization.
- Estimation of the parameters with K-means algorithm for ML method. This algorithm performs final foreground/background segmentation.
- Reconstruction of images color space and production of the restored manuscript.

### 3.1. Background Light Intensity Normalisation Algorithm

Background light intensity normalization algorithm is applied on historical manuscripts documents presented an uneven background and low contrast. Therefore, the choice of color space is important. This technique performs background approximation at first. Secondly, foreground normalisation is carried out from approximated background  $Y_{Back}$  and luminance image  $Y_{Original}$  of the original image. Figure 3 shows the normalization process.

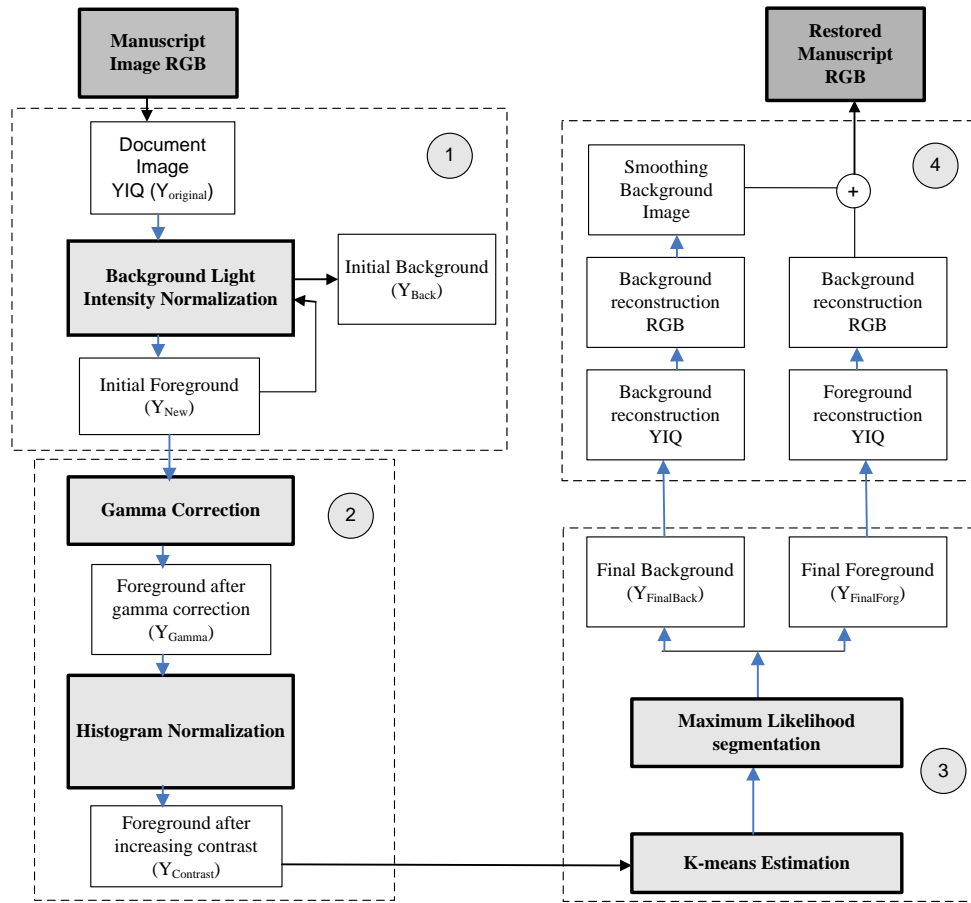


Figure 2: Flowchart of proposed methodology

### 3.1.1. Feature Choices

The choice of YIQ (Y: luminance channel; I and Q: chrominance color channels) colorspace is justified by the fact that the human vision is very sensitive to the change of luminosity. Moreover, the variation in light intensity caused by the uneven background of historical manuscripts is captured in L channel. An example of image decomposition from RGB to YIQ colorspace is presented in Figure 4.

### 3.1.2. Background Approximation

Background approximation algorithm start with a global binarization of the Y channel using Otsu method [Ots79]. This technique computes a global threshold for text extraction based on minimizing the intraclass variance of the image's pixels. The steps of the background approximation algorithm are presented below:

- Computing the horizontal projection profile H of binary document image from L channel.
- Computing the average of histogram values M (step1).

- Scanning the image line by line  $Y_{Original}$  and background approximation  $Y_{Back}$  (step 2).
- Recursive estimation of each final pixel grayscale of the image  $Y_{Back}$  (step3).

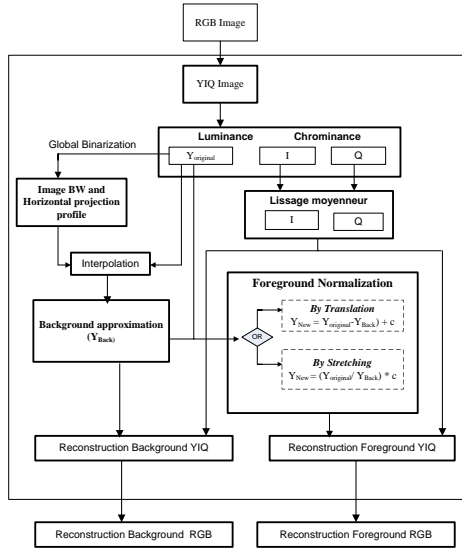
After experimental work, for the case of historical manuscripts, we suggest the following parameter values:  $window\ size = 3 \times 3$ ,  $mtime = 20$ . An example of a resulting background approximation is given in figure 4.

### 3.1.3. Image Normalisation

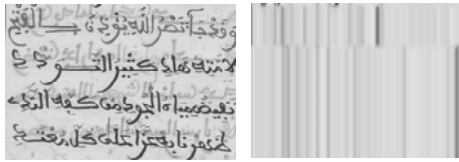
Foreground normalisation is obtained from L channel of original image and estimated background  $Y_{Back}$ . The light intensity pixels values of the new foreground  $Y_{New}$  are computed according to the following formulas, equation 1 and equation 2:

- Linear normalisation by translation

$$Y_{New} = (Y_{Original} - Y_{Back}) + C \quad (1)$$



**Figure 3:** Background light intensity normalisation process



**Figure 4:** Background light intensity normalisation on luminance channel  $Y$ : (left) Manuscript image, (right) Approximated background

- Linear normalisation by stretching

$$Y_{New} = \left( \frac{Y_{Original}}{Y_{Back}} \right) \cdot C \quad (2)$$

To ensure that the value does not exceed 255,  $C$  is set to 255 (usually) to make the background color white. The resultant normalized images are shown in Figure 5.

The normalisation by stretching is more adapted for the case of manuscript documents which present an uneven background and a low contrast. Experimental works show that iterating the normalisation process is necessary. We suggest three iterations which gives sufficient results. The obtained foreground  $Y_{New}$  is used in the later steps.

## 3.2. Histogram Transformations

### 3.2.1. Gamma Correction

Cheng and al, and Tremeau and al [CJSW01, TFMB04] have shown in their survey on color space, that the image processing with YIQ color space requires a gamma correction. In our case, the coefficient of gamma correction is the ratio between the average of intensity values of original image  $Y_{Original}$

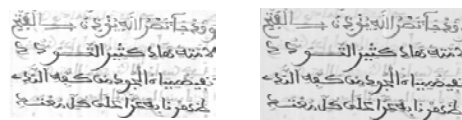


**Figure 5:** Foreground normalisation: (top)  $Y_{Original}$  image, (left) By translation, (right) By stretching

original and the resulting foreground  $Y_{New}$ , according to the following formula, equation 3.

$$\gamma = \frac{Mean(Y_{New})}{Mean(Y_{Original})} \quad (3)$$

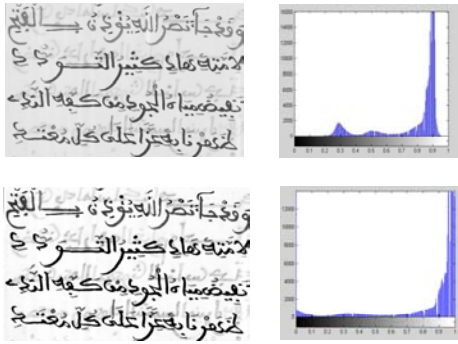
We notice that the values of gamma are usually greater than 1. This operation increases the contrast of image  $L_{New}$ . Figure 6 shows the effect of gamma correction.



**Figure 6:** Gamma correction: (left) Foreground before gamma correction, (right) Foreground after gamma correction

### 3.2.2. Histogram Normalisation

After gamma correction, the resulting foreground  $Y_{Gamma}$  contains again pale colors. In order to increase the contrast of image, we apply a stretching to the intensity values of image histogram using a proportion value. Then, the image  $Y_{Contrast}$  is produced with a proportion between 2% and 8% which gives correct results shown in figure 7.



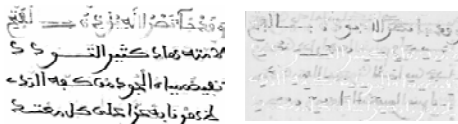
**Figure 7:** Histogram normalisation: (top) Foreground after gamma correction, with the intensity histogram, (bottom) Foreground after contrast adjustment, with the intensity histogram

### 3.3. Foreground/Background Segmentation

Document image manuscript segmentation can be considered as a statistical classification problem. The estimation of parameters of classification is given by Kmeans algorithm improved by ML method.

#### 3.3.1. Initialisation of K-means Algorithm

K-means algorithm operates on the image  $Y_{Contrast}$ . This technique computes the statistical features vectors for the foreground/background classification. K-means algorithm performs a first foreground/background classification. Figure 8 illustrates the result of segmentation.



**Figure 8:** K-means foreground/background segmentation: (left) Foreground, (right) Background

We notice that there is a significative loss of text information from foreground. In order to improve results of final segmentation, we refine the parameters of classification estimated by k-means algorithm by using ML algorithm.

#### 3.3.2. Maximum Likelihood Algorithm

The maximum likelihood classifier is one of the most popular methods of classification in remote sensing, in which a pixel with the maximum likelihood is classified into the corresponding class. The likelihood  $L_k$  is defined as the posterior probability of a pixel belonging to class K, and it computed as the following, equation 4 .

$$L_k = P(K|X) = \frac{P(K) \cdot P(X|K)}{\sum P(i) \cdot P(X|i)} \quad (4)$$

Where:

- $P(K)$ : prior probability of class k
- $P(X|K)$ : conditional probability to observe X from class k, or probability density function

Therefore  $L_k$  depends on  $P(X|k)$  or the probability density function. Foreground/background segmentation of  $Y_{contrast}$  image is performed by ML method. This technique relies on the likelihood function of the distribution of image intensity pixels. The ML method estimates the probability that a pixel belongs to its corresponding class which is foreground or background and assigns it when its probability is maximal. We are using two probability distributions, the Gaussian law and the Raleigh law. According to the distribution, the likelihood  $L_k$  is expressed in the following equations 5 and 6.

- $L_k$  according to Gaussian distribution:

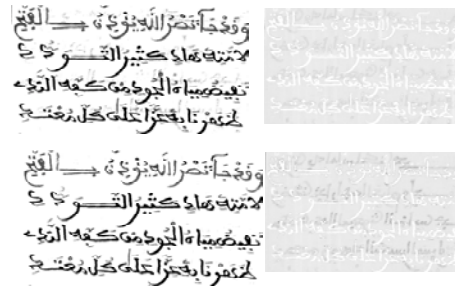
$$L_{k=1,2}(Y_{Contrast}) = \frac{1}{\sigma_k \sqrt{2\pi}} \cdot e^{-\frac{1}{2\sigma_k^2} \cdot (Y_{Contrast} - \mu_k)^2} \quad (5)$$

- $L_k$  according to Raleigh distribution:

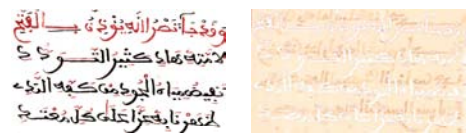
$$L_{k=1,2}(Y_{Contrast}) = \frac{1}{\mu_k \sqrt{\frac{2}{\pi}}} \cdot e^{-\frac{Y_{Contrast}^2}{2(\mu_k \sqrt{\frac{2}{\pi}})^2}} \quad (6)$$

The pixel  $j$  in  $Y_{Contrast}$  is labelled  $L_{k_j}$  according to the following equation 7.

$$L_{k_j} = \max_x (L_k(Y)) \quad (7)$$



**Figure 9:** ML Foreground background segmentation: (left) Gaussian distribution, (right) Raleigh distribution



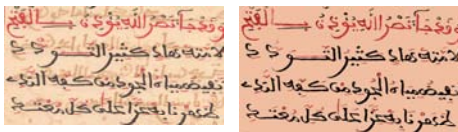
**Figure 10:** Foreground/background RGB Reconstruction: (left) Foreground, (right) Background



Experimental work shows that for the case of historical manuscripts, the Raleigh distribution gives better results for foreground/background segmentation. Figures 9 and 10 illustrate results in HSL and RGB color spaces.

### 3.4. Manuscripts Restoration

Foreground/background segmentation by ML method is used for the restoration of historical manuscripts. In fact, the restored image is constructed by superposition of the foreground and the average of background in RGB colorspace. Figure 11 illustrates visually the restored historical manuscript.



**Figure 11:** Manuscript output: (left) Original, (right) Restored

## 4. Conclusion and Future Works

In this paper, we have presented a hybrid method for foreground/background segmentation suited for Arabic documents manuscripts distorted with show-through effects and uneven background. This technique is based on four steps: (a) foreground extraction with iterative light intensity normalisation algorithm, (b) postprocessing of obtained foreground with double contrast adjustment, (c) Foreground/background segmentation with ML method, (d) reconstruction of restored manuscript document.

Our future objective aims to perfect our method. We are planning to assist the user to define automatically the iterations number of the normalisation process. Moreover, we aim to improve the texture segmentation method in order to classify the document into three types of texture information: Text, background, and graphic, can be useful to an indexing and retrieval system of Arabic historical documents.

### Acknowledgment

The Authors will Thanks Prof. Adel Alimi, Director of the Ecole Nationale des Ingénieurs de Sfax - ENIS, Head of the Laboratory REGIM. Special thanks to Prof. Abdellatif Benabdehaid, director of the GED (Equipe Gestion Électronique de Document), University of Le Havre. Special thanks to Dr. Volker M"argner, Head of the image processing group, at the Department of Signal Processing for Mobile Informations Systems , Institute for Communications Technology (IfN), Braunschweig Technical University.

## References

- [BHH\*98] BOTTOU L., HAFNER P., HOWARD P., SIMARD P., BENGIO Y., LECUN Y.: High quality document image compression with djvu. *Journal of Electronic Imaging* 7, 3 (July 1998), 410–425.
- [CJSW01] CHENG H., JIANG X., SUN Y., WANG J.: Color image segmentation: advances and prospects. In *Pattern Recognition* (2001), vol. 34, pp. 2259–2281.
- [CR00] C.A.B.MELLO, R.D.LINS: Image segmentation of historical documents. *Visual2000* (September 2000).
- [CR02] C.A.B.MELLO, R.D.LINS: Generation of images of historical documents by composition. *ACM Symposium on Document Engineering* (2002).
- [GPH05] GARAIN U., PAQUET T., HEUTTE L.: On foreground-background separation in low quality color document images. In *Eighth International Conference on Document Analysis and Recognition (ICDAR)* (2005), pp. 585–589.
- [KI86] KITTLER J., ILLINGWORTH J.: Minimum error thresholding. *Pattern Recognition* 19, 1 (1986), 41–47.
- [KSW85] KAPUR J. N., SAHOO P. K., WONG A. K. C.: A new method for gray-level picture thresholding using the entropy of the histogram. *Computer Vision, Graphics, and Image Processing* 29 (1985), 273–285.
- [LBE04] LEYDIER Y., BOURGEOIS F. L., EMPTOZ H.: Serialized k-means for adaptative color image segmentation: Application to document images and others. In *Document Analysis Systems (DAS)* (2004), pp. 252–263.
- [LVPG02] LEEDHAM G., VARMA S., PATANKAR A., GOVINDARAYU V.: Separating text and background in degraded document images - a comparison of global thresholding techniques for multi-stage thresholding. In *Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR'02)* (2002), p. 244.
- [LYT\*03] LEEDHAM G., YAN C., TAKRU K., TAN J. H. N., MIAN L.: Comparison of some thresholding algorithms for text/background segmentation in difficult document images. In *Seventh International Conference on Document Analysis and Recognition (ICDAR)* (2003), pp. 859–869.
- [Ots79] OTSU N.: A threshold selection method from gray level histogram. *IEEE Transactions in Systems, Man, and Cybernetics* 9 (1979), 62–66.
- [SG04] SHI Z., GOVINDARAJU V.: Historical document image enhancement using background light intensity normalization. *The 17th International Conference on Pattern Recognition (ICPR) 01* (2004), 473–476.
- [SG05] SHI Z., GOVINDARAJU V.: Historical document image segmentation using background light intensity normalization. *SPIE Electronic Imaging, Document Recognition and Retrieval XII 5676*, 1 (2005), 167–174.

- [TFMB04] TRÉMEAU A., FERNANDEZ-MALOIGNE C., BONTON P.: *Image numérique couleur de l'acquisition au traitement*. Dunod, Paris, 2004.
- [WXLT03] WANG Q., XIA T., LI L., TAN C. L.: Document image enhancement using directional wavelet. *Computer Vision and Pattern Recognition (CVPR) 02* (2003), 534.