Journal of Universal Computer Science, vol. 14, no. 2 (2008), 284-298
submitted: 15/6/07, accepted: 17/11/07, appeared: 28/1/08 © *J.UCS*

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

Wafa Boussellaa
(University of Sfax, REGIM, ENIS
Sfax, Tunisia
wafa.boussellaa@gmail.com)

Abderrazak Zahour
(University of Le Havre, GED Group, IUT
Le Havre, France
abderrazak.zahour@univ-lehavre.fr)

Adel Alimi
(University of Sfax,, REGIM, ENIS
Sfax, Tunisia
adel.alimi@ieee.org)

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.

Keywords: Segmentation, restoration, light intensity normalisation, k-means, maximum likelihood, foreground/background, Arabic historical color manuscript image.
Categories: I.4.3, I.4.4, I.4.5, I.4.9

1 Introduction

The historical documents, preserved at the National Library of Tunisia [BibNat], are considered as an important part of Arabic cultural heritage. These funds suffer from a progressive degradation and therefore risk disappearing. 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 In fact, historical documents have many particularities which hinder classical color document image segmentation algorithms.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.a), the presence of spot due to the humidity absorbed by paper (Figure 1.b), and an uneven background color paper (Figure 1.c).

(a) (b) (c)

Figure 1: Arabic historical documents image: (a) Show-through effects, (b) Localized spots, (c) uneven background.

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 [Otsu, 79], entropy techniques proposed by Kapur and al. [Kapur, 85] and the minimal error technique by Kittler and Illingworth [Kittler, 86], are analysed and compared in [Leedham, 02] and [Leedham, 03]. Another entropy-based method specially designed for historical document segmentation [Mello, 00] deals with the noise inherent in the paper especially in documents written on both sides. Wang and al. [Wang, 03] 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 [Mello, 02].

The works described in [Bottou, 98] and [Garain, 05] and [Leydier, 04] are dedicated for foreground/background separation of color document images. DjVu [Bottou, 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 [Leydier, 04] 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 [Garain, 06] have proposed an adaptive 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 [Bottou, 98].

Shi and al [Shi, 04] and [Shi, 05] 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

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.

2 Proposed Method

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 [Shi, 04] and [Shi, 05] 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.

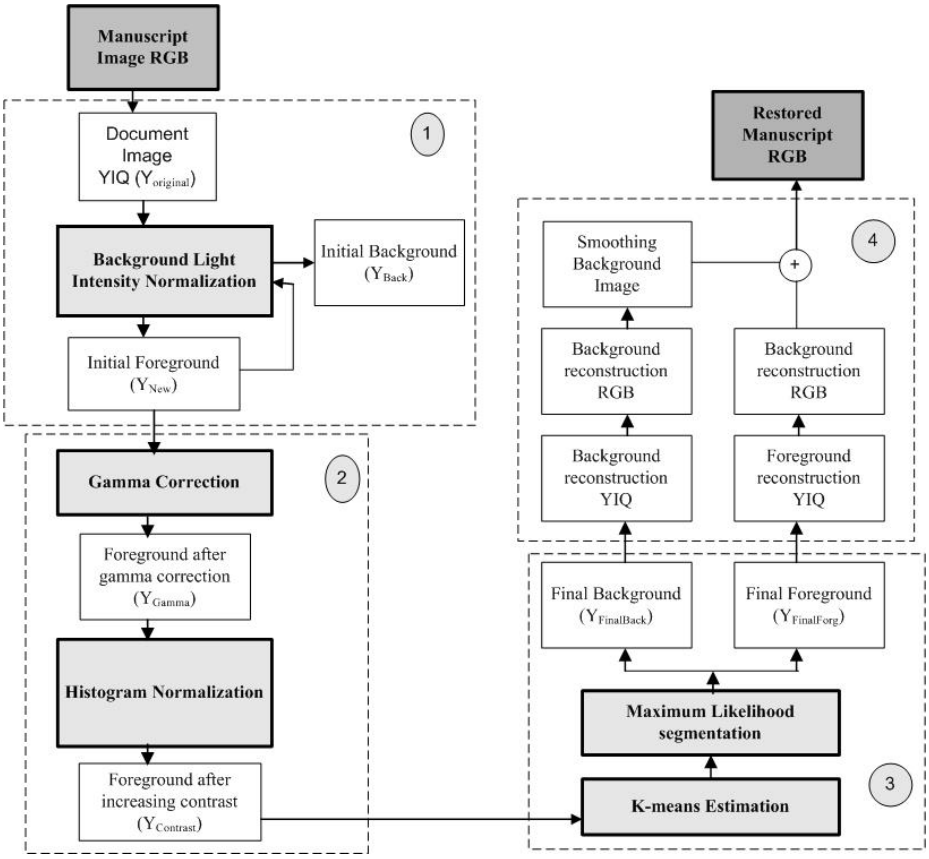


Figure 2: Flowchart of proposed methodology

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.
- Parameters estimation 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.

2.1 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.

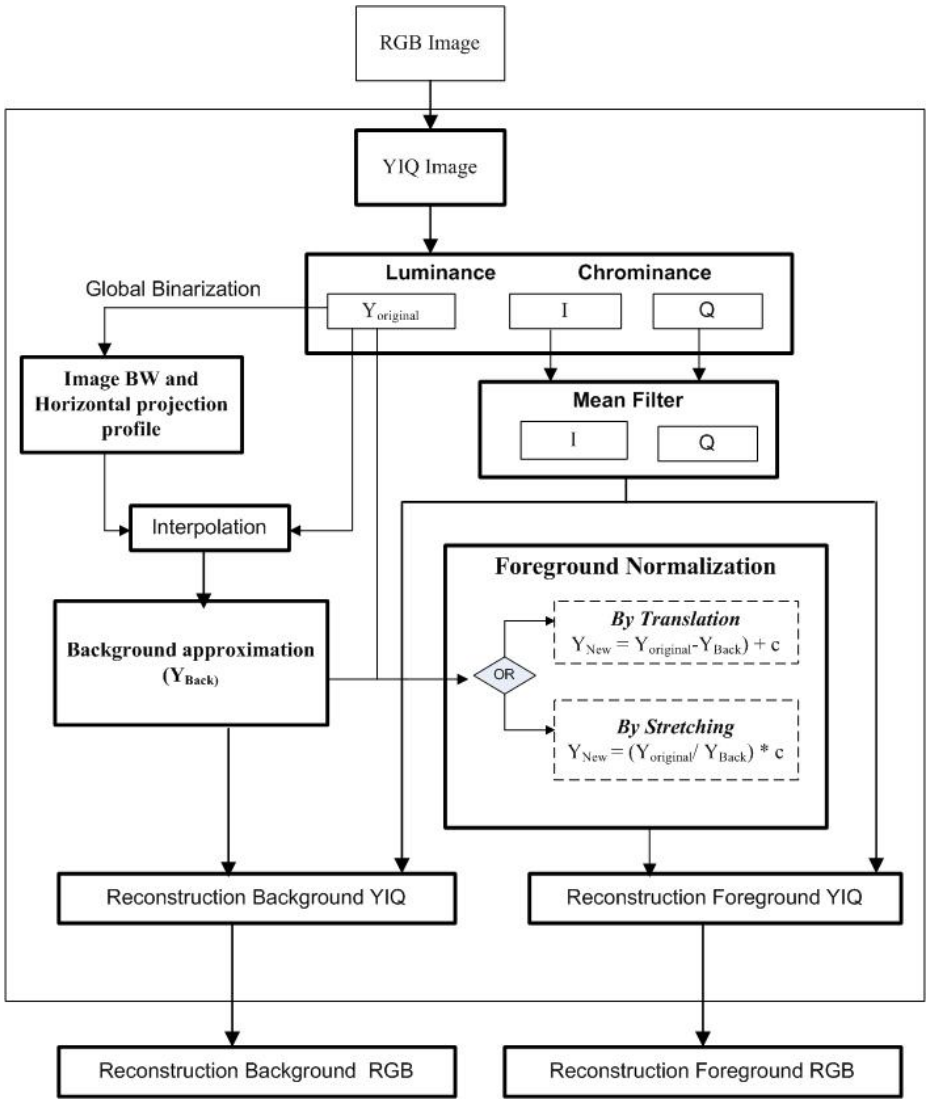


Figure 3: Flowchart of Background light intensity normalisation process

2.1.1 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 variation of luminosity. Moreover, the variation in light intensity caused by the uneven background of historical manuscripts is captured in Y channel. Compared with the HSV colorspace, the hue channel is less sensitive to the variation in light intensity and separates the objects having different colors. An example of image decomposition from RGB to YIQ and HSV colorspaces is presented in Figure 4.

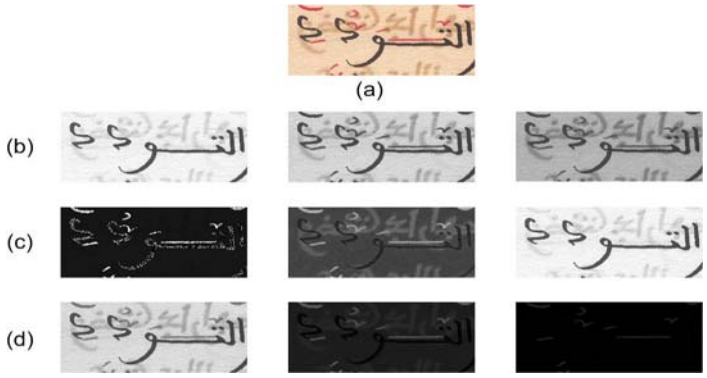


Figure 4: Colorspace transformations: (a) Original Image, (b) RGB colorspace, (c) HSV colorspace, (d) YIQ colorspace

2.1.2 Background Approximation

Background approximation algorithm start with a global binarization of the Y channel using Otsu's method [Otsu, 79]. This technique computes a global threshold for text extraction based on minimizing the intraclasses variance of the image's pixels.

The steps of the background approximation algorithm are presented below:

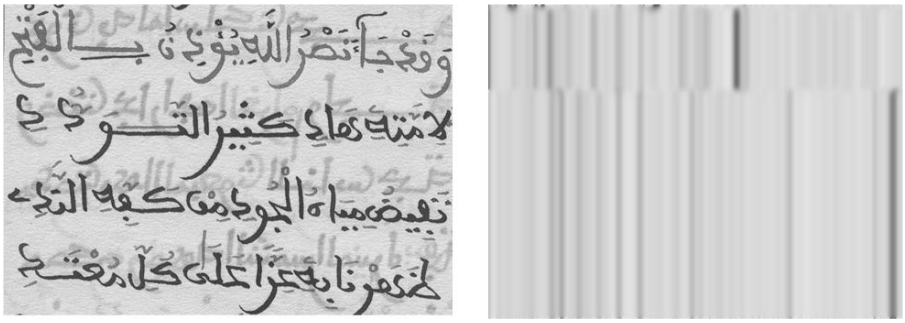
- Computing the horizontal projection profile H of binary document image from L channel.
- Computing the average of histogram values M .
- Scanning the image line by line $Y_{original}$ and background approximation Y_{Bak}..
- Recursive estimation of each final pixel's grayscale of the image Y_{Bak} with a an experimental fixed number of iterations (mtime) and a moving window size s×s .

Algorithm Background Approximation
Output : Y_{Back} (x, y): Approximated Background
Input : Y(x,y)**:**The L channel with size Xmax × Ymax
Initialisations : mtime : The number of iterations of background approximation with moving window; Initialized to 20. Output matrix initialized to 0.

1. M=mean(H)
2. For i= 0 to Xmax
   If H (i) < M
   $Back_{current}$ = Y(i)

$Back_{previous} = Back_{current}$
else
$Back_{current} = Back_{previous}$
endif
$Y_{Back} = Y_{Back} + Back_{current}$
End
3. count = 0
While (count < mtime)
- Scanning $Y_{Back}(x,y)$ with a moving window of size $n \times n$.
- Each pixel value is modified with the average of $n \times n$ pixels neighbours.
- count = count + 1
End
4. End

After experimental work, for the case of historical manuscripts, we suggest the following parameter values: window size = 3×3, mtime = 20. An example of a resulting background approximation is given in figure 5.



(a) (b)

Figure 5: Background light intensity normalisation on luminance channel Y: (left) Manuscript image, (right) Approximated background.

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.

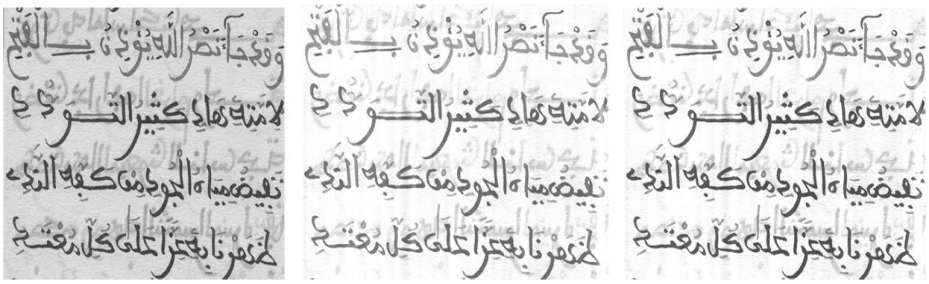
- Linear normalisation by translation

$$Y_{New} = (Y_{original} - Y_{Back}) + C \qquad (1)$$

- Linear normalisation by stretching

$$Y_{New} = (Y_{original} / Y_{Back}) * C \qquad (2)$$

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 6.



(a) (b) (c)

Figure 6: Foreground normalisation: (a) $Y_{Original}$ image, (b) By translation, (c) By stretching

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.

2.2 Histogram Transformations

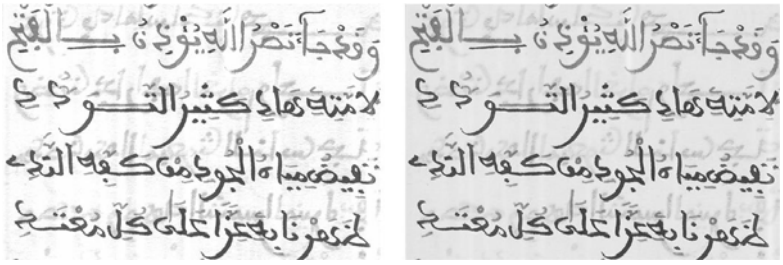
2.2.1 Gamma correction

Cheng and al [Cheng, 01] and Trémeau and al [Trémeau, 04] 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 $Y_{original}$ and the resulting foreground Y_{New}, according to the following formula, equation 3.

$$\gamma = \text{Mean}(Y_{New}) / \text{Mean}(Y_{original}) \quad (3)$$

We notice that the values of gamma are usually greater than 1. This operation increases the contrast of image L_{New}. Figure 7 shows the effect of gamma correction.



(a) (b)

Figure 7: Gamma correction: (a) Foreground before gamma correction, (b) Foreground after gamma correction.

2.2.2 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 8.

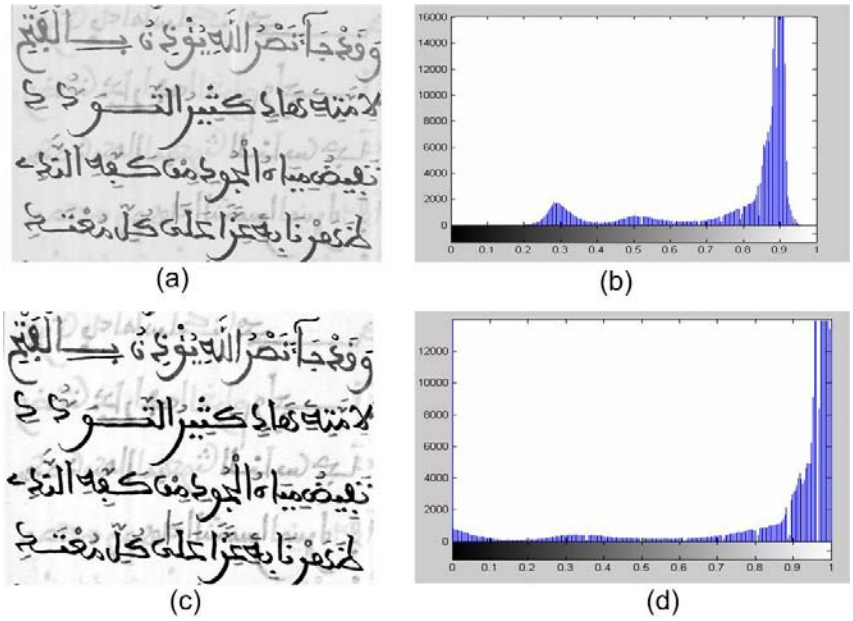


Figure 8: Histogram normalisation: (a) Foreground after gamma correction, (b) Intensity histogram (a), (c) Foreground after contrast adjustment, (d) Intensity histogram.

2.3 Foreground/Background Segmentation

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

2.3.1 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 is initialized as below.

Algorithm K-means

1. Partitioning the image $Y_{contrast}$ to M_n square windows.

n: The number of mask in image.

2. For each mask M_i we associate a feature vector: $\overline{x_i}(\mu_i, \sigma_i)$ With

$x_i^1 = \mu_i$: mean of intensity pixels' values of M_i.

$x_i^1 = \sigma_i$: Standard deviation of d of intensity pixels' values of M_i.

3. Initializing the centers of classes $\overline{c}_{k=1,2}$ where: $\overline{c_1} = \mathrm{Min}(\overline{x}_{i=1,\ldots,n})$; $\overline{c_2} = \mathrm{Max}(\overline{x}_{i=1,\ldots,n})$

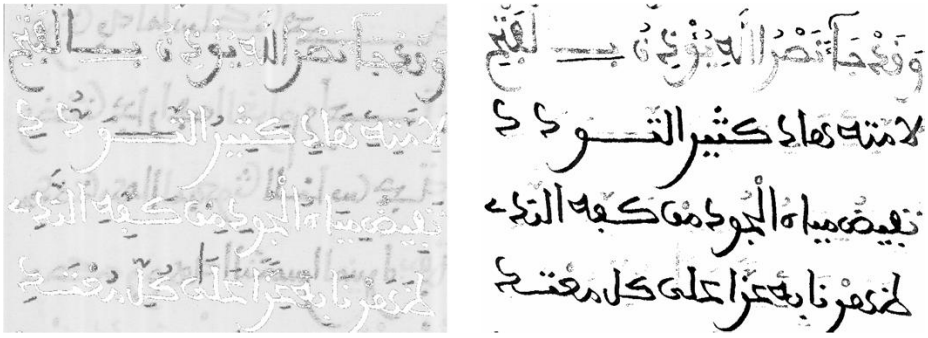
4. while E change

- For each vector $\overline{x}_{i=1..n}$
- Labelling of each mask M_i with the C_j while : $\overline{d}(\overline{x_i}, \overline{c_j}) <= \overline{d}(\overline{x_i}, \overline{c_k}) \forall k \neq j$
- For each class $C_{j=1,2}$, recalculate the average vector of centers:
- $\overline{c_j} = 1/|c_j| \sum_{M_{l \in c_j}} \overline{x_l}$
- Computing the error function:

$$E = \sum_{j=1}^{k(2)} \sum_{M_{l \in c_j}} \overline{d}(\overline{x_l} - \overline{c_j})$$

5. End

K-means algorithm performs a first foreground-background classification. Figure 9 illustrates the result of segmentation.



(a) (b)

Figure 9: K-means foreground-background segmentation: (a) Foreground, (b) Background

We notice that there is a significant 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.

2.3.2 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 Lk is defined as the posterior probability of a pixel belonging to class K, and it computed as the following, equation 4.

$$Lk = P(K/X) = (P(K)*P(X/K))/\sum P(i)*P(X/i) \qquad (4)$$

Where:

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

Therefore Lk depends on P(X|k) or the probability density function. Foreground/background segmentation of Ycontrast 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 Lk is expressed in the following equations 5 and 6.

- Lk according to Gaussian distribution:

$$L_{k=1,2}(Y_{contrast}) = \frac{1}{\sigma_k \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_k^2}(Y_{contrast} - \mu_k)^2\right) \qquad (5)$$

- Lk according to Raleigh distribution:

$$L_{k=1,2}(Y_{contrast}) = \frac{1}{\mu_k \sqrt{2/\pi}} \exp\left(-\frac{Y_{contrast}^2}{2(\mu_k \sqrt{2/\pi})^2}\right) \qquad (6)$$

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

$$L_{k_j} = \max_k (L_k(Y)) \qquad (7)$$

Experimental work shows that for the case of historical manuscripts, the Raleigh distribution gives better results for foreground-background segmentation. Figures 9 and 10 and 11 illustrate results in HSL and RGB color spaces.

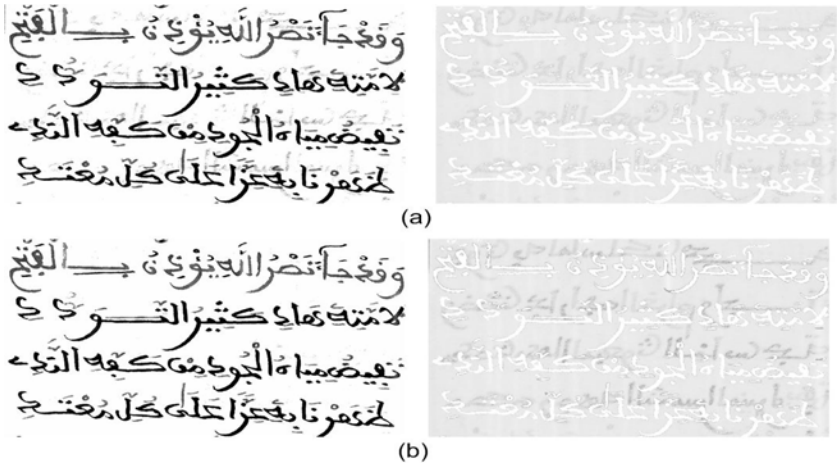


Figure 10: ML Foreground background segmentation: (top) Gaussian distribution, (bottom) Raleigh distribution.

Figure 11: Foreground/background RGB Reconstruction: (left) Foreground, (right) Background.

2.4 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 12 and 13 illustrates visually the restored historical manuscript.

(a) (b) (c)

Figure 12: Manuscript output: (left) Original, (right) Restored

Figure 13:Manuscript output: (left) Original, (right) Restored

Our method is tested on a set of Arabic historical manuscripts documents images from the National Tunisian Library. The images were scanned at 300 dpi and saved as TIF format without compression. Experimental work presented in Figures 14 and 15 shows that the produced results on manuscripts distorted with show-through effects, uneven background and low contrast are successful.

(a) (b) (c) (d)

Figure 14: Manuscript with show-through effects: (a) Original, (b) Foreground (c) Background, (d) Restored

(a) (b) (c) (d)

Figure 15: Manuscript with show-through effects: (a) Original, (b) Foreground (c) Background, (d) Restored

Otherwise, the tests achieved on documents images distorted with localized spot are shown in Figure 16. These results must be improved in our application.

(a) (b) (c) (d)

Figure 16: Manuscript with localized spot: (a) Original, (b) Foreground (c) Background, (d) Restored

3 Conclusion

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) post processing of obtained foreground with double contrast adjustment, (c) Foreground/background segmentation with ML method, (d) reconstruction of restored manuscript document.

Our proposed document image enhancement method process both color and grey scale document images. Test document set are selected from Arabic civilisation database. Our tests could be extended to Latin documents.

4 Future Work

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 segmentation process by using the EM algorithm and evaluating the performance of our foreground/background process behind one known foreground/background method of DJVU system [Bottou, 98].

Acknowledgements

The Authors Thanks Prof. Abdellattif Benabdehafid and Prof. Bruno Taconet from University of Le Havre for their encouragements and advices to accomplish this research. Special thanks to the National library of Tunisia [BibNat] for the access to their large document images database of Arabic historical documents.

References

[BibNat] La bibliothèque Nationale Tunisienne, http:// www.bibilotheque.nat.tn.

[Bottou, 98] L.Bottou, P.Haffner, and P.G Howard, High Quality Document Image Compression with DjVu, Journal of Electronic Imaging, 7(3), pp. 410-425, July 1998.

[Cheng, 01] H. D. Cheng, X. H. Jiang, Y. Sun, Jingli Wang, Color image segmentation: advances and prospects, Pattern Recognition 34 (2001) 2259-2281.

[Garain, 06] U. Garain, Thierry Paquet and Laurent Heutte, On Foreground-Background Separation in low Quality Document Images, International Journal of Document Analysis (2006) 8(1): 47–63.

[Kapur, 85] P. J.N.Kapur and A.K.C.Wong, A new method for gray-level picture thresholding using the entropy of the histogram, Computer Vision, Graphics, and Image Processing, vol. 29, pp.273-285, 1985.

[Kittler, 86] J.Kittler and J.Illingworth, Minimum error thresholding, Pattern Recognition, vol. 19, no. 1,pp. 41-47, 1986.

[Leedham, 02] G. Leedham, S. Varma, A. Patankar, and V. Govindaraju, Separating text and background in degraded document images - a comparison of global thresholding techniques for multi-stage thresholding, in Proceedings Eighth International Workshop on Frontiers of Handwriting Recognition, September 2002.

[Leedham, 03] G. Leedham, C. Yan, K. Takru, Joie H. Nata Tan and L. Mian, Comparison of Some Thresholding Algorithms for Text/Background Segmentation in Difficult Document Images, in Proceedings of the Seventh International Conference on Document Analysis and Recognition, Janvier 2003.

[Leydier, 04] Y. Leydier, F.L. Bourgeois, and H. Emptoz, Serialized K-Means for Adaptive Color Image Segmentation-Application to Document Images and Others, in 6th International Workshop on Document Analysis systems (DAS), Itay, LNCS vol.3163, 252-263, 2004.

[Mello, 00] C.A.B.Mello and R.D.Lins, Image segmentation of historical documents, in Visual2000, Mex-ico City, Mexico, September 2000.

[Mello, 02] C.A.B.Mello and R.D.Lins, Generation of images of historical documents by composition, in ACM Symposium on Document Engineering, McLean, VA, USA, 2002.

[Otsu, 79] N.Otsu, A threshold selection method from gray level histogram, IEEE Transactions in Systems, Man, and Cybernetics, vol. 9, pp. 62-66, 1979.

[Shi, 04] Z. Shi and V. Govindaraju, "Historical Document Image Enhancement Using Background Light Intensity Normalization". 17th International Conference on Pattern Recognition, Cambridge, United Kingdom, 23-26 August 2004.

[Shi, 05]Z. Shi and V. Govindaraju, Historical Document Image Segmentation Using Background Light Intensity Normalization, SPIE Document Recognition and Retrieval XII, San Jose, California, USA16-20 January 2005.

[Trémeau, 04] A. Trémeau, C. Fernandez-Maloigne, P. Bonton, Image numérique couleur de l'acquisition au traitement, ISBN 2 10 006843 1, Dunod, Paris, 2004.

[Wang, 03] Q. Wang, T. Xia, L. Li, and C. Tan, "Document image enhancement using directional wavelet," in Proceedings of the 2003 IEEE Conference on Computer Vision and Pattern Recognition, Madison, Wisconsin, USA, June 2003.