2009 10th International Conference on Document Analysis and Recognition

Enhanced Text Extraction from Arabic Degraded Document Images using EM Algorithm

Wafa Boussellaa[1] , Aymen Bougacha[1], Abderrazak Zahour[2], Haikal EL Abed[3], Adel Alimi [1]
[1]*University of Sfax, REGIM, ENIS, Route Soukra, BPW, 3038, Sfax, Tunisia*
[2]*IUT, Université du Havre, Place Robert Schuman, 76610 Le Havre, France*
[3]*Technical University Braunschweig, Institute for Communication Technology (IfN), Germany*
Email:{wafa.boussellaa,adel.alimi,elabed}@ieee.org, Aymen.Bougacha@gmail.com, abderrazak.zahour@univ-lehavre.fr

Abstract

This paper presents a new enhanced text extraction algorithm from degraded document images on the basis of the probabilistic models. The observed document image is considered as a mixture of Gaussian densities which represents the foreground and background document image components. The EM algorithm is introduced in order to estimate and improve the parameters of the mixtures of densities recursively. The initial parameters of the EM algorithm are estimated by the k-means clustering method. After the parameter estimation, the document image is partitioned into text and background classes by the means of ML approach. The performance of the proposed approach is evaluated on a variety of degraded documents comes from the collections of the National library of Tunisia.

1. Introduction

The automatic processing of degraded historical documents is a challenge in document image analysis field which is confronted with many difficulties due to the storage condition and the complexity of their content. For historical degraded and poor quality documents, enhancement is not an easy task. The main interest of an enhancement step of historical documents is to remove information coming from the background. Background artifacts can derive from many kinds of degradation, such as scan optical blur and noise, spots, underwriting, or overwriting. Most previous document image enhancement algorithms have been designed primarily for binarization. Binarization aim to extract text from distorted degraded documents and its related methods are proposed for processing gray documents which have not been extended for color documents.

In this paper an enhanced text extraction method is proposed on the basis of the maximum likelihood (ML) estimation for the segmentation problem. The difficult task lies in how to estimate the parameters of the likelihood functions and the number of segments. Eventually, the expectation maximization algorithm (EM) algorithm is introduced in order to improve the parameter estimation. The initial estimates for the EM algorithm are given by the k-means clustering algorithm to avoid the problem of random initial selection. The text and background segmentation is performed by the conventional ML method. The segmented image is used to produce a final colored restored document image

The rest of this paper is organized as follows. Section 2 gives an overview of previous work of degraded document image enhancement. Section 3 and 4 describes the proposed algorithm in details. Experimental results are presented in section 5. Conclusion and future work are given in last section.

2. Related work

According to the literature, approaches that deals with document image enhancement and text extraction are based on binarization or foreground/background separation techniques. Most previous document image enhancement algorithms have been designed primarily for binarization. Binarization is performed either globally or locally. Global thresholding methods are not sufficient since document images usually are degraded including shadows, non-uniform illumination and low contrast. Local methods are shown to perform better according to recent exhaustive survey of image binarization methods presented in [15]. Two main approaches are distinguished, based local thresholding methods and clustering based methods.

Based local thresholding techniques have been proposed to estimate a different threshold for each

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.220

pixel according to the grey-scale information of the neighboring pixels. Some of these popular methods, namely Otsu's thresholding technique [13] locally adaptive technique [11,14].
Other adaptive methods specially designed for historical and distorted documents are based on adaptive threshold segmentation. Gatos et al. [8] presents a binarization methodology based on background estimation used to segment the image, various pre- and post-processing steps are needs in this approach. Oh et al. [12] presents an iterative algorithm based on water flow models and a hierarchical thresholding. This method deals with low contrasted documents images.

An iterative approach for segmenting degraded document images is described by Kavallieratou et al [10]. It consist in obtaining an initial segmentation using a global thresholding and applying a local thresholding on the areas which are incorrectly segmented and detected.

Other methods for historical document image enhancement are driven by the goal of improving human readability of the documents are Based clustering methods. These methods are dedicated for foreground/background separation of color document images using a classification approach. Garain et al. [7] have proposed an adaptive method for foreground-background separation in low quality color document images. A connected component labeling 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 which representing local uniformity in illumination background.

Drira et al. [6] proposed a recursive method of unsupervised clustering. It classifies the pixels of document image in three clusters (background, original text, and show-through effect) in the degraded document. Thereafter the show-through effect must be eliminated and replaced by the color of the background.
Agam et al. [2] have described a novel approach based on probabilistic models (EM) for foreground and background separation. This algorithm deals with low contrasted document images.

3. The EM Algorithm

The EM algorithm mentioned in [5] is an iterative algorithm for calculating the maximum-likelihood or maximum-a-posterior estimates when the observations can be viewed as incomplete data. Each iteration of the algorithm consists of an expectation step followed by a maximization step.

The observed document image is considered as a mixture model of two Gaussian densities which represents the foreground (YF) and background (YB) document image components. The data of this model are determined by a random vector X of the probability density function (PDF) which is written as the following, equation 1:

$$F(x,\theta) = \sum_{k=1}^{K} \pi_k f_k(x;\theta_k) \qquad (1)$$

Where:

- K =2: The number of densities assumed a priori.
- $\theta_1 = (\mu_{YB}, \sigma_{YB})$: The mean and the standard deviation vectors of YB component.
- $\theta_2 = (\mu_{YF}, \sigma_{YF})$: The mean and the standard deviation vectors of YF component.

The PDF $f_1(x;\theta_1)$ and $f_2(x;\theta_2)$ are used for maximum likelihood clustering. Then, the estimation of θ_1 and θ_2 parameters are performed using of the EM algorithm which need an initializing parameter step. This step is the first one of our segmentation algorithm detailed in the following section.

4. Proposed Method

The proposed approach is considered as a novel view and improvement of our proposed methodology published in [3]. This approach belongs to our system PRAAD (Pre-processing and Analysis Tool for Arabic Ancient Document) [4].

Our approach process both color and grayscale document images. To apply our approach on color document image, we convert the image to YIQ colors space and operate on Y luminance channel. This choice 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 poor degraded is captured in Y channel.
According to the poor quality of document images and the pale colors and degradations coming from the background artifacts which affect the foreground contrast, we apply a stretching to the intensity values of image histogram using a proportion value. Then, the image YC is produced with proportion between 2% and 8% which gives correct results. After this needed pre-processing step, the contrasted image is operated by the segmentation algorithm detailed in the following section.

4.2. The segmentation algorithm

Our proposed text and background segmentation algorithm operates in three steps which are presented below.

4.2.1. Initial estimation

Initial estimates θ_1 and θ_2 and their corresponding mean and standard deviation vectors $(\mu_{YB}^{(0)}, \sigma_{YB}^{(0)}, \mu_{YF}^{(0)}, \sigma_{YF}^{(0)})$ for EM algorithm are calculated using k-means clustering method presented in [3].

4.2.2. Iterative Estimation by EM algorithm

The EM algorithm is iteratively carried out with the initial estimates $\theta^{(0)}$ and the intensity histogram H of YC document image. The EM algorithm converges when difference of old estimates and new estimates are less than some threshold ε and the final estimates $\theta^{(EM)}$ is obtained. The EM algorithm contributes to the segmentation algorithm by way of improving the parameters of the mixture of densities on the basis of the ML criterion. The EM algorithm in initialized as below.

Algorithm EM

Input:

- $K=2$
- $\theta^{(0)} = [\pi_1^{(0)}, \pi_2^{(0)}, \mu_1^{(0)}, \sigma_1^{(0)}, \mu_2^{(0)}, \sigma_2^{(0)}]$: estimates vectors by k-means algorithm.

Where:

  - $(\mu_1^{(0)}, \sigma_1^{(0)}, \mu_2^{(0)}, \sigma_2^{(0)})$: Means and standard deviation vectors
  - $(\pi_1^{(0)}, \pi_2^{(0)}) = (\frac{1}{2}, \frac{1}{2})$

- **H** a histogram vector defined previously
- ε the threshold for the algorithm convergence

Output:

$\hat{\theta}^{(0)}$: Local Maximum of likelihood law

$t \leftarrow 0$;

$\hat{\theta}^{(0)} = \theta^{(0)}$: Model initialization

Repeat

(Expectation step)

Compute the posterior probabilities $\hat{z}_{i,k}^{(t)}$.

$$\hat{z}_{i,k}^{(t)} = \frac{\hat{\pi}_k^{(t)} f(H_i; \hat{\theta}_k^{(t)})}{\sum_{l=1}^{K} \hat{\pi}_l^{(t)} f(H_i; \hat{\theta}_l^{(t)})}$$

(Maximization Step)

- Estimates of π_k, μ_k, σ_k maximizing $Q(\theta; \hat{\theta}^{(t)})$

$$\hat{\pi}_k^{(t+1)} = \frac{\sum_{i+1}^{n} \hat{z}_{i,k}^{(t)}}{n}$$: Estimates a priori probability $\hat{\pi}_k$ of k-th density of the mixture.

$$\hat{\mu}_k^{(t+1)} = \frac{\sum_{i=1}^{n} \hat{z}_{ik}^{(t)} H_i}{\sum_{i=1}^{n} \hat{z}_{ik}^{(t)}}$$: Estimates a priori means $\hat{\mu}_k$ of k-th density of the mixture.

$$\hat{\sigma}_k^{(t+1)} = \frac{\sum_{i+1}^{n} \hat{z}_{ik}^{(t)} (H_i - \hat{\mu}_k^{(t+1)})^2}{\sum_{i=1}^{n} \hat{z}_{ik}^{(t)}}$$: Estimates covariance matrix $\hat{\sigma}_k$ of k-th density of the mixture.

Until $\left\| Q(\theta; \hat{\theta}^{(t)}) - Q(\theta; \hat{\theta}^{(t-1)}) \right\| < \varepsilon$.

4.2.3. Maximum likelihood segmentation

The image segmentation is carried out by the conventional ML method using $\theta^{(EM)}$. The ML method estimates the probability that a pixel belongs to its corresponding class which is text or background and assigns it when its probability is maximal. We are using the probability distributions of Raleigh law. According to the distribution, the likelihood is expressed in the following equations (According to the Raleigh law):

$$f_{k=1,2}(YC) = \frac{1}{\hat{\mu}_k \sqrt{\frac{2}{\pi}}} \exp\left(-\frac{YC^2}{2(\hat{\mu}_k \sqrt{\frac{2}{\pi}})^2}\right)$$

The pixel (i,j) in YC is labeled $V_k(i,j)$ according to the following equation:

$$\begin{cases} V_k(i,j) = 1 & si \quad f_k(YC(i,j)) = \max(f_1(YC(i,j)), f_2(YC(i,j))) \\ V_k(i,j) = 0 & si \quad non \end{cases}$$

Experimental works presented in our last work presented in [3], proved that for the case of degraded manuscripts, the Raleigh distribution gives better results for text-background segmentation. Figure 1 and Figure 2 show segmentation results of YC image both in grayscale and color space.

5. Experimental Results

As mentioned in the related work study, the document images enhancement methods are driven by the goal of improving human readability of document text. To evaluate the performance of our proposed method in enhanced text extraction, we achieve our tests on 100 scanned images of old degraded handwritten documents given by the National library of Tunisia [1].

To assess the performance of our method especially for difficult cases, three degradation types were selected for evaluation and the images are visually inspected.
The proposed method is evaluated using two sets of metrics. The first sets are based on two selected and correlated binarization criteria. The criteria are misclassification error (ME) and the relative foreground area error (RAE) proposed by Sezgin and Sankur [15]. In order to be calculated, these criteria need the ground-truth binary image which is manually obtained. The expected values of the above criteria vary between [0, 1]. In all cases, the measure that is closer to zero corresponds to the best binarization result. The analytical score values of the two criteria obtained for three types of degradations document image are shown in Table 1.

The second sets used the precision and recall criteria presented in [16]. The two criteria are defined below and results are shown in Table 2.

Precision = *No. of correctly pixel's foreground extracted / Nbr of all pixel's foreground extracted by the proposed method*

Recall = *No. of correctly pixel's foreground extracted / total Nbr of pixel's foreground present in the document.*

Where the number of pixel's foreground present in the document is calculated by using the ground-truth image. Precision reflects the performance of removing the degradation and recall reflects the performance of extracting the foreground.

These evaluations include comparative results between the proposed method and some known binarization algorithms [8, 11, 12, 13, 14]. Moreover an average performance score is computed for global decision. The result in Table 1 and Table 2 shows respectively that our method achieves the best average value in binarization criteria and in precision and recall criteria.

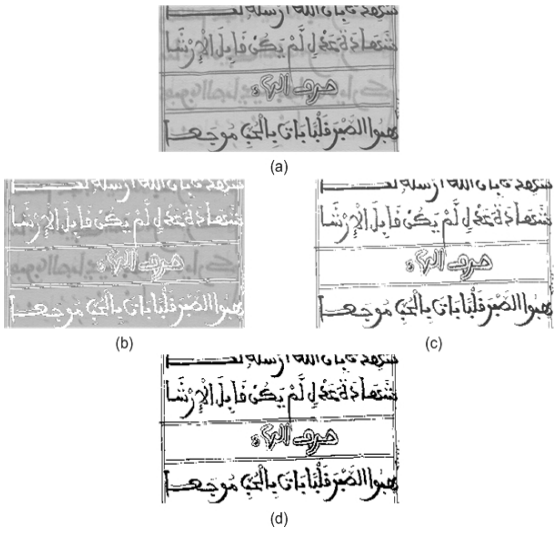

Figure 1. ML Text background segmentation in Y channel: (a) Original degraded image; (b) Background, (c) Text, (d) bitonal image.

Figure 2. ML Text background segmentation in RGB color space: (a) Original degraded image, (b) Background, (c) Text, (d) Reconstructed image.

Table 1. Evaluation' scores for two binarization criteria obtained by document distortion type

	Show-through effects		Localized spots		Uneven background.		AVE
	ME	RAE	ME	RAE	ME	RAE	
Niblack	0,114	0,119	0,089	0,086	0,160	0,168	0,122
Sauvola	0,035	0,038	0,061	0,066	0,060	0,060	0,053
Otsu	0,005	0,005	0,024	0,023	0,024	0,027	0,018
Gatos	0,047	0,035	0,032	0,034	0,042	0,041	0,038
Oh et al.	0,008	0,009	0,029	0,027	0,022	0,019	0,019
Proposed Method	0,003	0,003	0,023	0,000	0,014	0,009	0,009

Table 2. Evaluation' accuracy for precision and recall obtained by document distortion type

	Show-through effects		Localized spots		Uneven background		AVE	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Niblack	53%	96%	66%	94%	41%	95%	53%	95%
Sauvola	99%	72%	99%	63%	96%	49%	98%	61%
Otsu	96%	100%	89%	99%	82%	100%	89%	99%
Gatos et al.	75%	93%	98%	81%	96%	65%	89%	79%
Oh et al.	100%	93%	97%	84%	97%	83%	98%	86%
Proposed Method	100%	97%	93%	93%	96%	90%	96%	93%

6. Conclusion and future work

The developed enhanced text extraction algorithm operates firstly with a pre-processing step based on contrast adjustment of the document image. Then, this image is segmented into text and background components with the EM based segmentation algorithm. This algorithm is composed of three steps: (1) EM initializing, (2) EM estimation, (3) ML segmentation. According to the evaluated results, our method performs the best average values compared by the others methods. These values are about 0.009 mean errors in enhanced text extraction and accuracy about 96% rate for precision and 93% rate for recall.

7. Acknowledgement

This research is carried out within the framework of the DAAD project "In the way of information society" and the research cooperation projects between Tunisia and German. The Authors thanks the National library of Tunisia and the National Archives of Tunisia [1] for the access to its large document images database of Arabic historical documents.

References

[1] National library of Tunisia
*h*ttp://www.bibliotheque.nat.tn.

[2] G. Agam, G. Bal, G. Frieder, O. Frieder, "Degraded document image enhancement", *Proc. SPIE* 6500, pp. C1–11, 2007.

[3] W. Boussellaa, A. Zahour, and A. Alimi, "A methodology for the separation of foreground/background in Arabic historical manuscripts using hybrid methods", Journal of Universal Computer Science, 14(2):284–298, 2008.

[4] W. Boussellaa, A. Zhour, B. Taconet, A. Alimi, and A. Benabdelhafid, "PRAAD: Preprocessing and analysis tool for Arabic ancient documents", In 9th International Conference on Document Analysis and Recognition, vol. 2, pp. 1058–1062, 2007.

[5] A. P. Dempster, N. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", J. Royal Statistical Soc., Series B (Methodological), vol. 1, no. 39, pp. 1-38, 1977.

[6] F. Drira, F. Le Bourgeois, H. Emptoz, "Restoring Ink Bleed-Through Degraded Document Images Using a Recursive Unsupervised Classification Technique", Document Analysis Systems VII , Springer Berlin/Heidelberg, 2006, pp.38-49.

[7] U. Garain, T. Paquet, L. Heutte, "On Foreground-Background Separation in low Quality Document Images", International Journal of Document Analysis 8(1): pp. 47–63, (2006).

[8] B. Gatos, I. Pratikakis, S. J. Perantonis, "Adaptive degraded document image binarization", Pattern Recognition 39 (3), pp. 317–327, 2006.

[9] M. Junker, R. Hoch, and A. Dengel, "On the Evaluation of Document Analysis Components by Recall, Precision, and Accuracy", Proc. Fifth Int'l Conf. Document Analysis and Recognition, pp. 713-716,1999.

[10] E. Kavallieratou, E. Stamatatos, "Improving the Quality of Degraded Document Images", Second International Workshop on Document Image Analysis for Libraries, 2006.

[11] W. Niblack, "An introduction to digital image processing", Prentice-Hall, Englewood Cliffs, NJ, pp. 115–116, 1986.

[12] H.-H. Oh, K.-T. Lim, and S.-I. Chien, "An improved binarization algorithm based on a water flow model for document image with inhomogeneous backgrounds", Pattern Recognition, 38(12): pp.2612–2625, 2005.

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

[14] J. Sauvola, M. Pietikainen, "Adaptive document image binarization", Pattern Recognition 33(2), pp. 225–236, 2000.

[15] M. Sezgin, B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation", J. Electron. Imaging, 13(1), pp.146–165, 2004.

[16] C. L. Tan, R. Cao, P. Shen, "Restoration of Archival Documents Using a Wavelet Technique", IEEE Trans. Pattern Anal. Mach. Intell., 24(10), pp. 1399-1404, 2002.