EUSIPCO 2013 1569743917

BINARIZATION OF HISTORICAL DOCUMENTS USING SELF-LEARNING CLASSIFIER BASED ON K-MEANS AND SVM

Amina Djema and Youcef Chibani

Speech Communication and Signal Processing Laboratory
Faculty of Electronic and Computer Sciences
University of Sciences and Technology Houari Boumediene
USTHB, EL-Alia, B.P. 32, 16111, Algiers, Algeria
{adjema, ychibani}@usthb.dz

ABSTRACT

This article aims to present a new binarization method of degraded historical document images. The new algorithm combines K-Means classification with a classical binarization method to generate a pure learning set and a conflict class. We use SVM classifier to manage the conflict class in order to make the final binarization that classifies each pixel of image document as foreground or background. Experiments are conducted on the standard datasets Dibco 2009 and Dibco 2011. The obtained results are very promising that allows opening a large margin of investigation.

Index Terms—Historical documents, binarization, classification, K-means, SVM.

1. INTRODUCTION

Historical documents are a rich part of our culture, civilization, science, archives and records. They represent our passage in this world. However, they show, in the most of time, some degradations that make image document unreadable and make character recognition software and document processing inefficient. Hence, various binarization methods have been developed in the last past decades [1].

Binarization is an important step of document image analysis. It allows extracting foreground information from the image and results an information coded either 0 for foreground (text, figure, table…) or 1 for background. It is an initial and a critical task for document processing because their results affect significantly subsequent process for some applications as image document analysis, word spotting and character recognition software [1].

In general way, binarization methods are separated in two main categories: direct binarization and indirect binarization [2]. The first one is based on gray level analysis and studies their properties to establish a binarization threshold. Comparison between the value of the gray level of the pixel and the threshold defines the pixel cluster. In the related work, the choice of the threshold can be conducted into two main ways: global and local [3].

The global thresholding uses a single threshold to binarize the image, which is calculated using global information of the image [3]. The value of each pixel of the image is compared to this only threshold. Hence, the pixel is classified as a foreground if its value is greater than the threshold, otherwise it is classified as background [3-4]. The most popular method is Otsu's method, which consists to separate the histogram of the image into two parts that makes the variance intra-class minimized [5]. The success of the method is based on the assumption that the histogram of image document is bimodal. But, this is rarely the case for historical documents [6].

Unlike global thresholding, local binarization techniques propose to generate a new threshold for each pixel or each region [6-7]. The threshold value is calculated from their neighborhood. For instance, Niblack [8] calculates the threshold using the mean and standard deviation of the neighborhood. This method allows extracting correctly the foreground but generates some noises in background region. Sauvola reduces the amount of noise generated in Niblack's method by applying the assumption that foreground and background have a gray scale level close to 0 and 255, respectively [9].

The second category of binarization method is based on classification which allows clustering and selecting features in order to make binarization [2-10]. The drawback of these methods is that they need a training set to improve the binarization performances and their results depend mainly on the quality of the training set [2].

However, historical document images present several irregular degradations that make getting a strong training set a challenging problem. Hence, we propose to build an own training set for each considered image. The training set and a conflict class is created from confronting traditional methods of clustering. The conflict class categorizes the pixels that are not well classified by traditional methods which is managed by a learning process based on Support

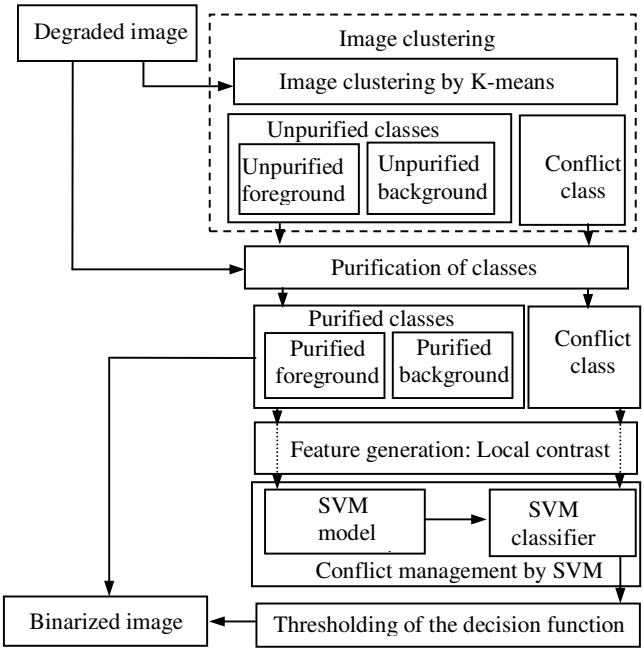


Fig. 1. The flowchart of the proposed binarization method.

Vector Machines (SVM).

The organization of the remaining paper is presented as follows. In section 2, we describe our method which is based on the combination of K-means classifier, Sauvola's method and SVM. In section 3, we present and discuss the results and we conclude in the last section in which we summarize our results and we propose future works.

2. METHODOLOGY

The proposed methodology for binarizing the degraded historical documents is divided into four steps as illustrated in Figure 1. The first step aims to generate three separate classes using K-means classification. In the second step, we consider a direct binarization in order to confront their result to classes generated in the previous step. This is aiming to prepare a strength learning set. The third step is a critical step where the pixel is classified as foreground or background using a supervised clustering SVM method. This classifier generates real decision function values. Thus, to get binary results, we need a threshold rule. Then, choosing an optimal threshold becomes a challenging problem which is discussed in the last step.

2.1. Image clustering

We consider the assumption that the image document has three classes: foreground, background and conflict [11-12]. To generate three distinct classes, we use K-means clustering with K=3 on the luminance component of the image in order to preclassify the pixel. This algorithm tries to separate the pixel of degraded image into K clusters [13]:

First, we choose randomly K initial means then we apply two iterative steps:

- Assignment step: Each feature vector extracted from the degraded image is assigned to the cluster which contains the closest centroid.
- Updating step: the means are calculated on the new assignment of the clusters.

These two steps are repeated until no changing is happened in assignement step.

2.2. Purification of classes

The conflict class will be processed by a supervised SVM classifier which requires a training set: foreground and background classes. However, these two classes contain no pure pixels. Therefore, it is important to enhance the purity of classes. Thus, we introduce a direct binarization.

In this work, we use Sauvola's method which calculates a local adaptive threshold $t(i,j)$ for each pixel (i,j) of the image document using the following formula [9]:

$$t(i,j)=\mu(i,j)\times\left(1-k\left(1-\frac{\sigma(i,j)}{R}\right)\right) \qquad (1)$$

Where $\mu(i,j)$ and $\sigma(i,j)$ are the mean and the standard deviation, respectively, calculated in the small window centered in the pixel (i,j). k is a positive constant taking the value in the range [0.2, 0.5]. R is the maximal value of the standard deviation.

A pixel is considered as pure if it does not change class value through the binarization methods. Let *Fg* and *Bg* refer to foreground and background classes, respectively. If the pixel (i,j) is assigned to *Fg* class by the K-means classifier and Sauvola's method, it is considered as a pure *Fg*, otherwise it is added to the conflict class. In the same way, if the pixel (i,j) is assigned to *Bg* class by the K-means classifier and Sauvola's method, it is considered as pure *Bg* , otherwise, it is added to the conflict class. This step allows generating a pure training set and avoiding to train classifier on inaccurate data.

2.3. Conflict management

When using K-means classifier and then Sauvola's method, some pixels are assigned to the conflict class. In order to separate between foreground and background classes, we propose to use a supervised classifier based on the Support Vector Machines (SVM).

SVM is introduced in 1995 by Vladmir Vapnik and immediately got success in many applications as pattern recognition [15]. The goal of the SVM is to divide data set into two classes according to data samples. This technique is

based on computation of the hyperplane that separates linearly the data set. The optimum separating hyperplane is the one that maximize the distance from both classes. The support vectors are training data that are closest from the hyperplane [16]. When training data are non-linearly separable, a kernel function is introduced in order to project training data in feature space which makes possible to find a linear separation between classes. Finally, the decision function is defined as:

$$f(x) = \sum_{m=1}^{N} \alpha_m y_m K(x, x_m) \quad (2)$$

Where α_m are the lagrangian multipliers, N is the number of training data, $K(x, x_m)$ is the kernel function and $y_m \in \{-1, +1\}$. A new data is classified referring to their position through the hyperplane. In our context, the SVM is used to classify the confusion pixel that usual binarization methods failed to classify it. In order to separate foreground and background, we use the local contrast as the feature vector for training the SVM.

2.4. Thresholding of the decision function

The optimum hyperplane can be represented by a decision function *f*, which is based on the sign of the decision function as shown by the following formula:

$$B(i,j) = \begin{cases} Fg & if\ f(x_{ij}) \geq 0 \\ Bg & otherwise \end{cases} \quad (3)$$

When the decision function takes a positive value, the pixel is assigned to *Fg* class, otherwise to *Bg* class. In this general case, the decision function compares its value to a null threshold, which can be considered as a hard thresholding. In order to get a more flexible use of the decision function, we consider a non-zero threshold. Then, the choice of the threshold becomes a challenging problem. Thus, we propose to adjust automatically the threshold in order to calculate its optimal value accounting two different approaches:

- The optimal threshold is adjusted according to the equal error rate (EER) which corresponds to the rate of the background error classification equal to the rate of foreground classification.
- The optimal threshold is adjusted according to the minimal error rate (MER) which corresponds to the sum of the foreground and the background error classification.

3. EXPERIMENTAL RESULTS

We carried out quantitative and visual criteria to evaluate our method, which is evaluated on two datasets: Document Image Binarization Contest (DIBCO) 2009 dataset which includes 5 printed historical documents and 5 handwritten historical documents while DIBCO 2011 dataset provides 8 printed historical documents and 8 handwritten historical documents. These images suffer from several degradations. To evaluate the performances of a binarization method, groundtruth images also are provided as reference binarized images. In order to evaluate the performances of our method according to the type of degradations, we define four types of degradation which are: Type 1: Ink bleed through, Type 2: Non-uniform background, Type 3: Smear and Type 4: Ink intensity variation. The evaluation is conducted on each type of degradation taking into account the kind of the document (handwritten, printed).

The most popular quantitative evaluations criterion is based on the F-Measure, which is defined as [17]:

$$F - Measure = \frac{2 \times recall \times precision}{recall + precision} \quad (4)$$

Such as:

$$recall = \frac{TP}{TP + FN} \quad (5)$$

$$precision = \frac{TP}{TP + FP} \quad (6)$$

TP refers to true positive foreground that means that the pixel (i, j) is classified as foreground either by the binarization method or the reference binarized image. *FP* refers to false positive foreground that means that the pixel (i, j) is classified as foreground by the binarization method but as background by the reference binarized image. *FN* refers to false negative background that means that the pixel is set as background by the binarization method unlike the reference binarized image where the pixel is classified as foreground.

Results are depicted in tables 1 and 2 for handwritten and printed document images, respectively and table 3 summarizes all evaluation conducted on two datasets.

As we can see, the proposed method based on MER outperforms K-means and Sauvola's binarization method for handwritten documents without taking into account the type of degradation. We notice that Sauvola's method deals better only with handwritten documents that suffer from smear and the proposed method yields better results for the image that suffers from the other degradations.

Table 2 demonstrates the performance of the K-means algorithm to solve binarization problem for printed documents. We notice, also, that Sauvola's method outperforms too for printed documents that suffer from smear.

Table 3 shows that, in general way, K-means deals

TABLE I. Comparative evaluation for handwritten documents based on F-measure.

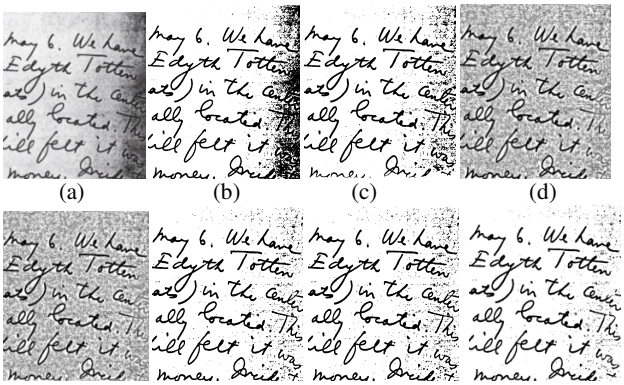
Degradation	K-means	Sauvola	Proposed threshold		
			Hard	*EER*	*MER*
Type 1	0.8410	0.7287	0.7843	0.7873	0.8462
Type 2	0.8109	0.8307	0.8969	0.8659	0.8764
Type 3	0.5546	0.8420	0.8260	0.8297	0.8247
Type 4	0.7616	0.8375	0.8540	0.8266	0.8404
Overall	0.7420	0.8097	0.8403	0.8273	0.8469

TABLE II. Comparative evaluation for printed documents based on F-measure.

Degradation	K-means	Sauvola	Proposed threshold		
			Hard	*EER*	*MER*
Type 1	0.8897	0.8809	0.8536	0.8550	0.8509
Type 2	0.8892	0.8360	0.7683	0.8094	0.8423
Type 3	0.8696	0.8882	0.7691	0.7905	0.8557
Type 4	0.9184	0.8201	0.7740	0.8571	0.8366
Overall	0.8917	0.8563	0.7912	0.8280	0.8463

TABLE III. Comparative evaluation for handwritten and printed documents based on F-measure.

Degradation	K-means	Sauvola	Proposed threshold		
			Hard	*EER*	*MER*
Type 1	0.8676	0.8178	0.8226	0.8255	0.8443
Type 2	0.8612	0.8286	0.7929	0.8179	0.8480
Type 3	0.7121	0.8651	0.7975	0.8101	0.8402
Type 4	0.8266	0.8302	0.8207	0.8393	0.8388
Overall	0.8081	0.8317	0.8098	0.8262	0.8418



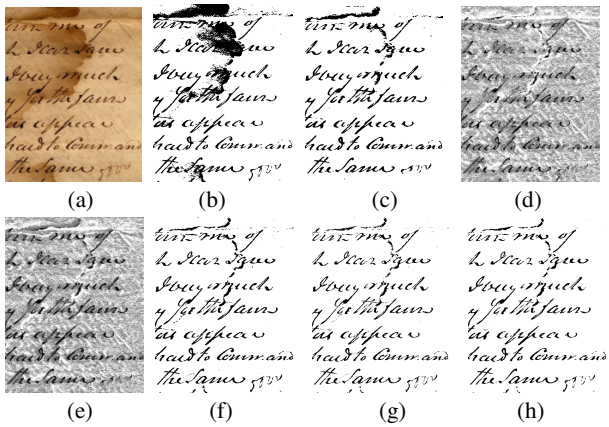
(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 2. Binarization of handwritten document having non-uniform background degradation: a) Original image, b) K-means method, c) Sauvola method, d) K-means classification, e) Purified classes, f) Hard threshold, g) Optimal threshold (EER) and h) Optimal threshold (MER).

better with image suffering from degradations type 1 and type 2 while the proposed method based on EER outperforms when image contains degradation type 4. Table 3 shows, also, that the proposed method with MER outperforms for the overall DIBCO dataset regardless the type of degradations.

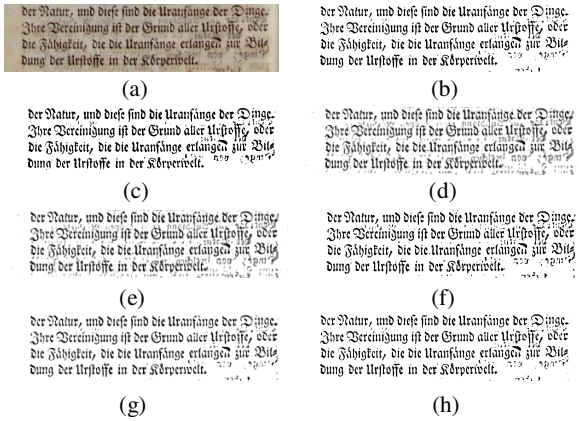
We can notice that the proposed method has provided, practically, the same performance whatever the type of degradations. This constitutes an additional advantage comparatively to K-means and Sauvola's method, which their performances depend on the type of degradation.

However, the quantitative measure doesn't represent



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 3. Binarization of handwritten document having ink intensity variation degradation: a) Original image, b) K-means method, c) Sauvola method, d) K-means classification, e) Purified classes, f) Hard threshold, g) Optimal threshold (EER) and h) Optimal threshold (MER).



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 4. Binarization of handwritten document having ink bleed through degradation: a) Original image, b) K-means method, c) Sauvola method, d) K-means classification, e) Purified classes, f) Hard threshold, g) Optimal threshold (EER) and h) Optimal threshold (MER).

significantly the performance of binarization method. This is why we appreciate the quality of binarization with visual criterion, too. Figures 2, 3, 4 and 5 show the results obtained for various selected thresholds. We note that the images indexed (e) for each figure have three intensities: the black and white intensity constitute the learning set which is purified. However, this image shows that a few number of pixels still misclassified. We note that a gray intensity represents a conflict class and we can see that they represent most of cases the degradation contained in the images. So, a correct classification of the conflict class should enhance the quality of binarization.

Figure 2 shows binarization results for handwritten document suffering from foreground intensity variation. We notice that all the method tested present some noises but our method yields better results and allows a decrease of the

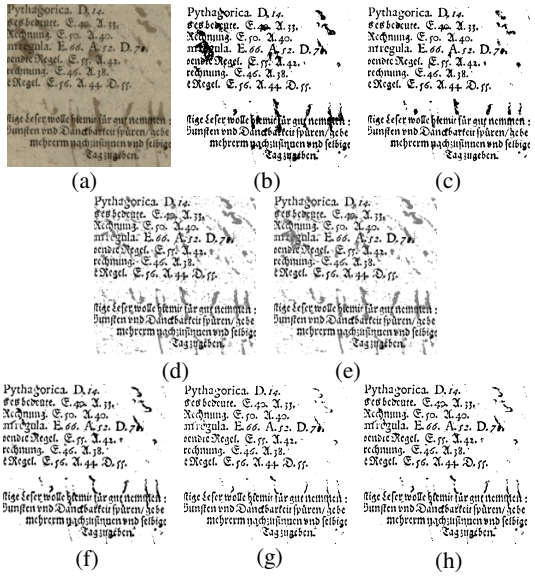


Fig. 5. Binarization of handwritten document having smear degradation: a) Original image, b) K-means method, c) Sauvola method, d) K-means classification , e) Purified classes, f) Hard threshold, g) Optimal threshold (EER) and h) Optimal threshold (MER).

amount of noise mainly for the one based on MER.

Figure 3 shows an example of binarization methods on image having ink intensity variation. We observe that our method based on MER and the one based on EER yield better foreground extraction than the other methods, mainly for the words "command" and "5000" contained into the image. This image contains an important smear. We observe that all the methods tested deal well excepting in the neighborhood edge of the smear and at this point we note that our method attempts better to eliminate the noise generated by the high local contrast.

Figure 4 shows results of binarization methods on printed document having ink bleed through degradation. We note that ours and K-means method have comparable results. Figure 4 (e) shows that the class of confusion contains all degradations and thus binarization error generated by our method is, mainly, due to bad training of the classifier.

Figure 5 shows results of binarization method on printed document having smear degradation. This figure shows the effectiveness of our method and mainly the one based on EER to decrease the impact of smear unlike Sauvola's method and K-means binarization where the smear and the text are confused.

4. CONCLUSION

This article presents a new approach to solve binarization problem. Our method combines two binarization methods for resolving the conflict between foreground and background classes using supervised classifier SVM. The first results are encouraging and promising.

This work constitutes a first step of investigation for the binarization of historical document according the type of degradation. Also, we attempt to investigate other methods of feature generation for training the SVM classifier in order to enhance performances of the proposed binarization method whatever the type of degradation.

5. REFERENCES

[1] Ø. D. Trier and T. Taxt, "Evaluation of binarization methods for document images," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 17, pp. 312-315, 1995.

[2] M. Valizadeh and E. Kabir, "Binarization of degraded document image based on feature space partitioning and classification," *International Journal on Document Analysis and Recogniti,*vol. 15, pp. 57-69, 2010.

[3] P. W. Palumbo, P. Swaminathan and S. N. Srihari, "Document Image Binarization: Evaluation of algorithms," *SPIE proceeding on applications of digital image processing IX*, vol. 697, pp. 278-285, 1986.

[4] R. Chamchong, C. C. Fung and K. W. Wong, "Comparing binarization techniques for the processing of ancient manuscripts," *International Federation for Information Processing*, Springer, Berlin, pp. 55–64, 2010.

[5] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Tran. Syst. Man Cybernet*, vol. 9, pp. 62-66, 1979.

[6] B. Gatos, I. Pratikakis and S. J. Perantonis, "An adaptive binarization technique for low quality historical documents," *Lecture Notes in Computer Science, Heidelberg: Springer,* vol. 3163, pp. 102-113, 2004.

[7] R. H. Singh, S. Roy, O. I. Singh, T. Sinam and Kh. M. Singh, "A new local adaptive thresholding technique in binarization," *International Journal of Computer Science issues*, vol. 8, pp. 271- 277, 2011.

[8] W. Niblack, An Introduction to Digital Image Processing. Prentice Hall int., 1986.

[9] J. Sauvola and M.Pietikainen, "Adaptive Document Image Binarization," *Pattern Recognition*, vol. 33, pp. 225-236, 2000.

[10] T. Sari, A. Kefali and H. Bahi, "An MLP for binarizing images of old manuscripts," *In Proc of IEEE International Conference on Frontiers in Handwriting Recognition*, pp. 247–251, 2012.

[11] B. Su, S. Lu and C. L. Tan, "A self-training learning document binarization framework1," *In proc of IEEE International Conference on Pattern Recognitio,* pp. 3187-3190, 2010.

[12] F. Drira, "Toward Restoring Historical Document Degraded Over Time," *In proc of IEEE International Conference on Document Analysis and Recognition*, pp. 350-357, 2006.

[13] A. Likas, N. Vlassis and J. J. Verbeek, "The global K-means clustering algorithm," *Pattern Recognition Societ*. vol. 36, pp. 451-461, 2003.

[14] T. Wakahara and K.Kita, "Binarization of color character strings in scene images using K-means clustering and support vector machines," *In proc of IEEE International Conference on Document Analysis and Recognition*, pp. 274-278, 2011.

[15] S. Marinai and H. Fujisawa, *Machines Learning in Document Analysis and Recognition*, Springer-Verlag, Berlin, 2008.

[16] N. V. Vapnik, *The Nature Of Statistical Learning Theory*, Springer, New York, 1995.

[17] B. Gatos, K. Ntirogiannis and I. Pratikakis, "ICDAR 2009 document image binarization contest (Dibco 2009)," *In proc of IEEE ICDAR*, pp. 1375-1382, 2009.