A new hybrid binarization method based on Kmeans

Mahmoud Soua, Rostom Kachouri, Mohamed Akil

A NEW HYBRID BINARIZATION METHOD BASED ON KMEANS

Mahmoud Soua *Rostom Kachouri* *Mohamed Akil*

soua.mahmoud@esiee.fr rostom.kachouri@esiee.fr mohamed.akil@esiee.fr

ESIEE Paris, LIGM, A3SI, 2 Bd Blaise Pascal, BP 99, 93162 Noisy-Le-Grand, France.

ABSTRACT

The document binarization is a fundamental processing step toward Optical Character Recognition (OCR). It aims to separate the foreground text from the document background. In this article, we propose a novel binarization technique combining local and global approaches using the clustering algorithm Kmeans. The proposed Hybrid Binarization, based on Kmeans (HBK), performs a robust binarization on scanned documents. According to several experiments, we demonstrate that the HBK method improves the binarization quality while minimizing the amount of distortion. Moreover, it outperforms several well-known state of the art methods in the OCR evaluation.

Index Terms— Scanned documents, binarization, Kmeans, OCR

1. INTRODUCTION

Document binarization is an important pre-processing for image analysis especially Optical Character Recognition (OCR) [1]. It tries to separate the foreground text from the background. In general, approaches that deal with binarization are categorized in two main classes: global and local [2]. Principally, global methods are based on histogram, classification and clustering approaches. In the histogram category, Rosenfeld and Kak [3] select one threshold from the histogram of 2D document. They assume that gray values of each object are located around each histogram peak. Another well-known method is the Iterative Global Thresholding (IGT) [4]. It is composed of two phases executed iteratively. Firstly, the average color of the image is computed and subtracted from the image. Then, a histogram stretching is performed, in such way that the remaining pixels take up all of the gray-scale tones. The main drawback of histogram methods is the inaccurate detection of peaks caused by noise [5]. In the classification approaches, Otsu [6] is a very popular technique [7]. It finds an automatic threshold that reduces the interclass variance between foreground and background. Many clustering methods are reported in the binarization litterature. The Fuzzy C-means (FCM) [8] is one of the most efficient algorithms [2]. It sets the image pixels into the correspondent foreground or background cluster. The membership degree of each pixel is computed according to a relatively complex fashion. Less complex [9], the Kmeans algorithm [10] clusters pixels according to a distance computation. According to the litterature, Kmeans is a simple and flexible clustering method [11] that can be employed in binarization [2]. In general, global binarization methods are efficient when they are applied to documents with uniform illumination. However, degraded documents, including stains, and uneven contrast, are not well processed by global methods [12]. To overcome this problem, local methods are used to binarize degraded documents. For example, Niblack [13] computes a threshold based on the mean and the standard deviation of small neighborhood around each pixel. Actually, the Niblack method identifies the text regions correctly as foreground, however, it tends to produce a large amount of noise in non-text regions. Sauvola [14] improves Niblack by using the dynamic range of image gray-value standard deviation. In case of low separation between foreground and background, the result is degraded significantly. Wolf [15] outperforms Sauvola by maximizing the local contrast. However, performance degradation arises if there is a sharp change in background gray values across the image. Recently, Sauvola Ms_{kx} [16] was proposed as an improvement to Sauvola. It begins with data sub-sampling. Then, binarization is performed at different scales. It should recognize that, local methods provide an adaptive solution since the binarization decision varies according to the properties of each document area. Local methods may introduce some noise like artefacts due to the local areas treatments. To improve the binarization quality, some recent works combine local and global approaches [2]. They perform an efficient local binarization while reducing the noise as the global one do [17]. For example, Gabara [18] uses local thresholding technique based on global edge detection. Hybrid IGT (HIGT) [19] applies the IGT global binarization [4] on the whole image then, reapplies it only on noisy areas. In most cases, the hybrid methods improve binarization quality on degraded documents.

In this context, we propose in this paper an adaptive method which we term Hybrid binarization based on Kmeans (HBK) to binarize scanned documents. Our method performs the Kmeans algorithm on local areas to get an improved re-

sult. Then, it employs a global technique to enhance local binarization. Performed experiments show that HBK is a robust binarization method. The obtained binarization allow to achieve higher OCR quality. The organization of the paper is shown as follow. Firstly, the Kmeans algorithm is presented in Section 2. We describe our proposed HBK method in Section 3. Next, in Section 4, experimental results are discussed. Finally, conclusion is drawn in Section 5.

2. KMEANS CLUSTERING

Kmeans [10] is one of the most popular unsupervised clustering algorithms, mainly used for data mining [20] and image processing [21]. Regarding the image processing context, Kmeans aims to partition all pixels in the image into k distinct clusters. Each cluster is represented by a single centroid.

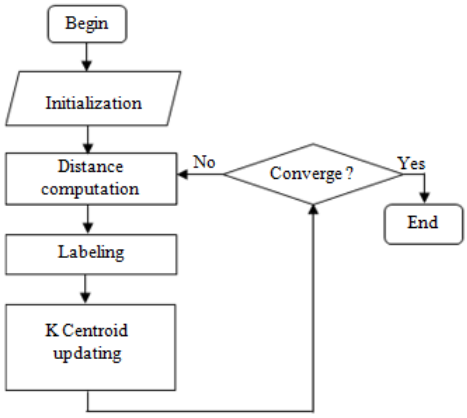

Fig. 1. The process flow of Kmeans

The Kmeans process flow is shown in Figure 1 and described as follow. In the first stage, the centroids number and values are initialized depending on the application. We note by $C = \{C_0, C_1, ..., C_i, ..., C_{k-1}\}$ the set of k input centroids. Afterward, the distances between centroids and pixels are computed to measure their similarity. For better clustering quality, the Euclidean distance is generally employed [22]. Given a centroid C_i and a pixel P_j the related Euclidean distance is defined as:

$$d(C_i, P_j) = \sqrt{(C_i^R - P_j^R)^2 + (C_i^G - P_j^G)^2 + (C_i^B - P_j^B)^2} \quad (1)$$

Where C_i and P_j are represented in the RGB color space by: $C_i = (C_i^R, C_i^G, C_i^B)^T$ and $P_j = (P_j^R, P_j^G, P_j^B)^T$, knowing that $i \in [0, k-1]$, $j \in [0, Np-1]$ and Np is the total number of pixels in the image. Next, the labeling stage sets the membership of pixels to the closest centroids according to the minimum computed distances. In addition, the pixel values and number in each cluster i are added respectively to the corresponding accumulator $Accu = \{Accu_0, Accu_1, ..., Accu_i, ..., Accu_{k-1}\}$, and counter $Count = \{Count_0, Count_1, ..., Count_i, ..., Count_{k-1}\}$ variable sets, with $Accu_i = (Accu_i^R, Accu_i^G, Accu_i^B)^T$. Finally, in the k centroid updating stage, each centroid C_i, $i \in [0, k-1]$ is re-estimated by computing the average assigned pixels to the corresponding cluster i. This process is done based-on the previous computed accumulators and counters. The distance computation, the labeling and the centroid updating stages are repeated until overall pixels are firmly assigned and centroid are no longer being changed. Before each iteration, accumulators and counters are reset to zero. In next section, we describe our proposal which is based fundamentally on the Kmeans algorithm.

3. PROPOSED METHOD : HYBRID BINARIZATION BASED ON KMEANS (HBK)

As a hybrid approach, our proposed binarization method performs higher local binarization robustness and reduces the possible appearing artifacts by using global correctness. In the rest of the paper we call this method HBK. Global and local phases of the HBK method are shown in Figure 2.

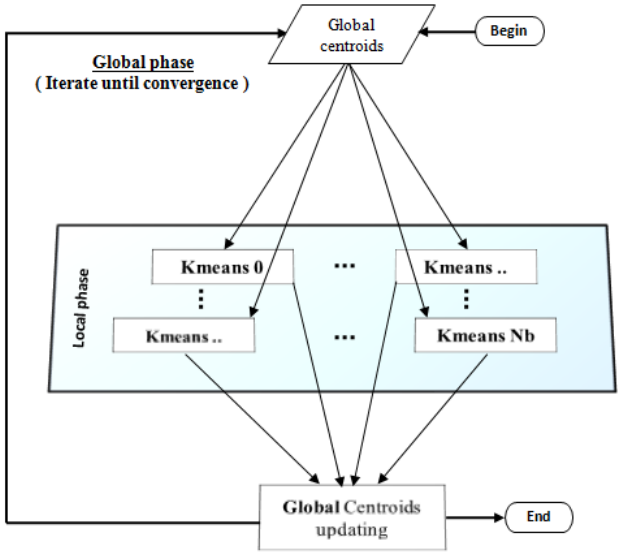

Fig. 2. The HBK method: global and local phases

Firstly, in the local phase, the input image (scanned document) is divided into Nb blocks. Then, the Kmeans algorithm is applied in each block. Afterward, the global phase gathers the obtained Kmeans results from each block iteratively and performs a correctness loop until convergence. More details of the HBK method are described in Algorithm 1.

Considering the binarization context, two initial centroids are employed. We note by $C_G = \{C_{G0}, C_{G1}\}$ the global centroid set. We initialize C_{G0} and C_{G1} with black and white

Algorithm 1 HBK(Img)

1: $C_{G0} \leftarrow (0,0,0)^T, C_{G1} \leftarrow (255,255,255)^T$
2: **while** $C_{Gi}(ite) \neq C_{Gi}(ite-1), i \in [0,1]$ **do**
3: **for** $bl \in [0, Nb-1]$ **do**
4: $Kmeans(bl, C_G)$
5: $Reset(Accu_G, Count_G)$
6: **for** $i \in [0,1]$ **do**
7: $Accu_{Gi} \leftarrow Accu_{Gi} + Accu_{Lbl,i}$ Eq.2
8: $Count_{Gi} \leftarrow Count_{Gi} + Count_{Lbl,i}$
9: **end for**
10: **end for**
11: $C_{Gi} \leftarrow \frac{Accu_{Gi}}{Count_{Gi}}, i \in [0,1]$ Eq.3
12: **end while**

values. Thus, $C_{G0} = (0,0,0)^T$ and $C_{G1} = (255,255,255)^T$ design respectively the background and the foreground clusters. The document is divided into blocks of equal sizes. Each block includes two local centroids given the set $C_L = \{(C_{L0,0}, C_{L0,1}), (C_{L1,0}, C_{L1,1}),\ ...,\ C_{Lbl,i},\ ...,\ (C_{LNb-1,0}, C_{LNb-1,1})\}$, with $C_{Lbl,i} = (C^R_{Lbl,i}, C^G_{Lbl,i}, C^B_{Lbl,i})^T$ given that $i \in [0,1]$, $bl \in [0, Nb-1]$. In the beginning, each local centroid is initialized with the corresponding global one: $C_{Lbl,0} = C_{G0}$ and $C_{Lbl,1} = C_{G1}, bl \in [0, Nb-1]$. In addition, each block has two local accumulators and counters given respectively by the sets $Accu_L$ and $Count_L$:

$$Accu_L = \{(Accu_{L0,0}, Accu_{L0,1}), (Accu_{L1,0}, Accu_{L1,1}),$$
$$..., Accu_{Lbl,i}, ..., (Accu_{LNb-1,0}, Accu_{LNb-1,1})\} and$$

$$Count_L = \{(Count_{L0,0}, Count_{L0,1}), (Count_{L1,0},$$
$$Count_{L1,1}), ..., Count_{Lbl,i}, ..., (Count_{LNb-1,0},$$
$$Count_{LNb-1,1})\}$$

with $Accu_{Lbl,i} = (Accu^R_{Lbl,i}, Accu^G_{Lbl,i}, Accu^B_{Lbl,i})^T$ where $i \in [0,1]$ and $bl \in [0, Nb-1]$.

As stated in line 4, the Kmeans algorithm is applied across all blocks in the scanned document. When the Kmeans process on each block converges, all the local accumulators are gathered into a global one: $Accu_G = \{Accu_{G0}, Accu_{G1}\}$ with, $Accu_{Gi} = (Accu^R_{Gi}, Accu^G_{Gi}, Accu^B_{Gi})^T, i \in [0,1]$. Line 7 states for the global accumulators compute. This process is performed by one addition per each color component as stated in Eq.2:

$$Accu_{Gi} = \begin{pmatrix} Accu^R_{Gi} + Accu^R_{Lbl,i} \\ Accu^G_{Gi} + Accu^G_{Lbl,i} \\ Accu^B_{Gi} + Accu^B_{Lbl,i} \end{pmatrix}, i \in [0,1], \quad (2)$$

$bl \in [0, Nb-1]$

Similarly, local counters of each block are accumulated into global ones given by $Count_G = \{Count_{G0}, Count_{G1}\}$. Finally, global accumulators $Accu_G$ are devided by the correspondent counters $Count_G$ to compute the C_{Gi} new centroids as shown in Eq.3:

$$C_{Gi} = \begin{pmatrix} C^R_{Gi} \\ C^G_{Gi} \\ C^B_{Gi} \end{pmatrix} = \begin{pmatrix} \frac{Accu^R_{Gi}}{Count_{Gi}} \\ \frac{Accu^G_{Gi}}{Count_{Gi}} \\ \frac{Accu^B_{Gi}}{Count_{Gi}} \end{pmatrix}, i \in [0,1] \quad (3)$$

If the global convergence is reached, the HBK algorithm stops, else it reiterates and local centroids of each block are reset with the computed global ones C_{G0} and C_{G1}. In the next section, we demonstrate that our method gives satisfying binarization quality thanks to the local approach.

4. EXPERIMENTAL RESULTS

In the following subsections, we present the employed dataset. Then, we demonstrate the efficiency of our proposed HBK method while reducing the distortion criterion. Finally, we compare our proposal to several state of the art methods in term of OCR evaluation.

4.1. The LRDE Document Binarization Dataset (LRDE-DBD)

In our experiments, we evaluate our HBK method on the LRDE-DBD[1] database [16]. It is composed of French text documents extracted from "Le Nouvel Observateur [2]" magazine. The provided dataset is composed of images with A4 format and 300-dpi resolution. Evaluation tests were performed on 125 scanned documents.

4.2. Distortion criterion-based evaluation

In general, to evaluate the efficiency of Kmeans clustering, the sum of cluster distortions is usually employed as a performance indicator [23, 24]. Given an image with Np pixels, we note by $\mathscr{C}(p_i)$ the Kmeans associated centroid of one pixel p_i, with $i \in [0, Np-1]$. The distortion criterion measure is defined by Eq.4:

$$distortion = \frac{1}{Np} \sum_{i=0} |p_i - \mathscr{C}(p_i)|^2 \quad (4)$$

[1]Copyright(c) 2012. EPITA and Development Laboratory (LRDE) with permission from Le Nouvel Observateur. LRDE-DBD is available online on the web site: $http://www.lrde.epita.fr/cgi-bin/twiki/view/Olena/DatasetDBD$

[2]Le Nouvel Observateur. Issue 2402, November 18-24, 2010 and available on the website: $http://tempsreel.nouvelobs.com$

For comparison reason, we implement a simple Kmeans binarization method that we term Global Binarization Kmeans (GBK). In this method, a single Kmeans is performed over the whole image. Keeping the goal of the binarization context, initial centroids of both GBK and HBK are set to the RGB values $C_{G0} = (0, 0, 0)^T$ and $C_{G1} = (255, 255, 255)^T$. To evaluate the clustering performances of these two methods, we measure the distortion criterion of GBK and HBK binarization on the LRDE-DBD database.

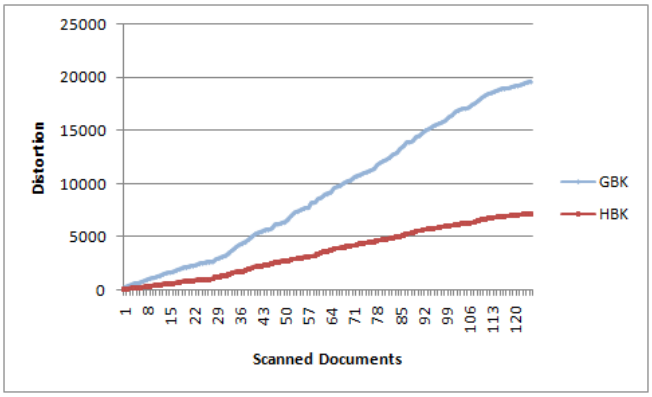

Fig. 3. Accumulated distortion computed from the obtained binarization results of GBK and HBK methods

Figure 3 shows the accumulated distortions of the two compared binarization methods on 125 scanned documents. The obtained result demonstrates that the HBK improves nearly 3x the distortion measure compared to GBK. Indeed, the treatment of Kmeans on small areas, gives local optimal distortions that minimizes the global one. We can come to the conclusion that our approach improves the binarization quality while reducing distortion in employed Kmeans clustering. In the next section, we perform an OCR evaluation to measure the HBK binarization efficiency.

4.3. OCR-based evaluation

Optical Character Recognition (OCR) is a process by which text characters contained in an image can be recognized. The binarization quality has a direct influence on the OCR result. In this context, the Tesseract 3.02 OCR [25] is used. Actually, we compare the character recognition rate of our HBK approach against seven well-ranked binarization methods [16, 14, 15, 6, 13, 26, 27] on the LRDE-DBD documents. Table 1 shows the OCR recognition rate of the evaluated methods. The first observation to be noted from this table, is that the OCR results are very close: The recent Sauvola Ms_{kx} gives an acceptable OCR rate by scoring 89% of accuracy. It ensures well text binarization but in some document areas, artifacts may appear leading to character miss-recognition. In the other side, the proposed HBK method gives the best OCR rate scoring 91% of accuracy thanks to its proper binarization results. Indeed, in the HBK method, the local binarization gives robust binarization quality and the global approach eliminates the artifacts generated by the local approach.

Table 1. OCR Accuracy evaluation of HBK and seven well-ranked binarization methods.

Methods	Images	OCR Accuracy (%)
HBK	Scanned	91
Sauvola Ms_{kx} [16]	Scanned	89
Wolf [15]	Scanned	88
Sauvola [14]	Scanned	87
Lelore [26]	Scanned	85
Otsu [6]	Scanned	84
Niblack [13]	Scanned	80
TMMS [27]	Scanned	73

5. CONCLUSION

Binarization is an important pre-processing step for optical character recognition (OCR). In this paper, we proposed a new hybrid binarization method based on the Kmeans clustering algorithm. We show that HBK performs robust binarization while minimizing the amount of distortion. Several experiments are performed on real magazine documents. The proposed HBK method outperforms several state of the art methods in the OCR-based evaluation while scoring 91% of accuracy. In future work, we intend to adapt the block sizes with kind of text in order to ensure multiscale binarization. In the other side, we note that the HBK method is highly adapted to parallel processing thanks to the no-dependency between image blocks. For this, we plan to use a parallel architecture in order to accelerate the HBK execution time.

6. REFERENCES

[1] P. Xiu, H.S. Baird, *Whole-Book Recognition*, Pattern Analysis and Machine Intelligence, IEEE Trans. vol. 34 , no. 12, 2012.

[2] P. Stathis, E. Kavallieratou, N.Papamarkos, *"An evaluation technique for binarization algorithms"*, Journal of Universal Computer Science, vol. 14, no. 18,3011-3030, 2008.

[3] A. Rosenfeld, and A. C. Kak, *Digital Picture Processing*, 2nd ed. New York: Academic, 1982.

[4] E. Kavallieratou, *"A binarization algorithm specialized on document images and photos"*, 8th Int. Conf. on document Analysis and Recognition, pp. 463-467, 2005.

[5] C. A. Glasbey, *"An Analysis of histogram-based thresholding algorithm, Graph"*, Models Image Processing 55, 532-537, 1993.

[6] N. Otsu, *"A threshold selection method from gray-level histograms"*, IEEE Transactions on Systems, Man and Cybernetics, 9(1): 62-66, 1979.

[7] F. Martn, *"Analysis tools for gray level histograms"*, Proc. Of SPPRA, pp. 11-16, 2003.

[8] R. Duda, and P. Hart, *"Pattern Classification and Scene Analysis"*, Wiley, New York, 1986.

[9] S. Ghosh, S.K. Dubey, *"Comparative Analysis of K-Means and Fuzzy C-Means Algorithms"*, (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 4, no. 4, 2013.

[10] S. P. LLOYD, *"Least square quantization in PCM"*, IEEE Transactions on Information Theory vol.28, no.2, 129-137, 1982.

[11] V. Pritesh , O. Bhavesh , "A Survey on K-mean Clustering and Particle Swarm Optimization", International Journal of Science and Modern Engineering (IJISME), 24-26, 2013.

[12] M. Valizadeh, E. Kabir, *"An adaptive water flow model for binarization of degraded document images"*, International Journal on Document Analysis and Recognition (IJDAR)June 2013, vol. 16, no. 2, pp 165-176, 2013.

[13] W. Niblack, *An Introduction to Digital Image Processing*, Standberg Publishing Company, 1985.

[14] J. Sauvola, M. Pietikainen, *"Adaptive document image binarization"*, Pattern Recognition, vol. 33, 225-236, 2000.

[15] C. Wolf, J.-M. Jolion, *"Extraction and recognition of artificial text in multimedia documents"*, Pattern Analysis and Applications, vol. 6, 309-326, 2004.

[16] G. Lazzara, T. Graud, *"Efficient Multiscale Sauvolas Binarization"*, International Journal on document analysis and recognition, 2013.

[17] M. P. Greenleaf, J. Shen, D. Yoon, *"Local-Global Image Binarization reconstructing the cellular structure of polymer foam materials"*, Computer-Aided Design and Applications, 10(6), pp 919-928, 2013.

[18] E. Gabarra, A. Tabbone, *"Combining global and local threshold to binarize document of images"*, Pattern Recognition and Image Analysis, vol: 3523, 173-186, 2005.

[19] V. Sokratis, E. Kavallieratou, R. Paredes, K. Sotiropoulos, *"A hybrid binarization technique for document images"*, Learning Structure and schemas from documents studies in computational intelligence. Vol. 375, 165-179, 2011.

[20] K. Singh, D. Malik, N. Sharma, *"Evolving limitations in Kmeans algorithm in data mining and their removal"*, IJCEM International Journal of Computational Engineering and Management, vol. 12,105-109, 2011.

[21] N.A.M. Isa, S.A. Salamah, *"Adaptive fuzzy moving K-means clustering algorithm for image segmentation"*, Consumer Electronics, IEEE Transactions on, vol. 55, no. 4, 2009.

[22] S. Cha, *"Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions"*, International Journal of Mathematical Models and Methods in Applied Sciences, 1(4),300-307, 2007.

[23] M. B. Al-Daoud, N. B. Venkateswarlu and S. A. Roberts, *"New methods for the initialization of clusters"*, Pattern Recognition Lett., vol. 17, no. 15, 451455, 1996.

[24] J. M. Pena, J. A. Lazano, P. and Larranaga, *"An empirical comparison of four initialization methods for the K-means algorithm"*, Pattern Recognition Lett., 20, 1027-1040, 1999.

[25] R. Smith, *"An Overview of the Tesseract OCR Engine"*, Proc. Ninth Int. Conference on Document Analysis and Recognition (ICDAR), IEEE Computer Society, 629-633 , 2007.

[26] T. Lelore, F. Bouchara, *"Super-resolved binarization of text based on the FAIR algorithm"*, Proceedings of international conference on Document Analysis and Recognition, 839-843, 2011.

[27] J. Fabrizio, B. Marcotegui, M. Cord, : *"Text segmentation in natural scenes using toggle-mapping"*. Image Processing (ICIP), 16th IEEE International Conference, 2349-2352, 2009.