Segmentation of complex document

Souad Oudjemia, Zohra Ameur
Department of Electrical Engineering,
University of Mouloud Mammeri, Algeria
E-mail: souadoudjemia@yahoo.fr

Abdeldjali Ouahabi
Signal & image Group,
University of François Rabelais, Tours
E-mail: abdeldjalil.ouahabi@univ-tours.fr

***Abstract*—In this paper we present a method for segmentation of documents image with complex structure. This technique based on GLCM (Grey Level Co-occurrence Matrix) used to segment this type of document in three regions namely, 'graphics', 'background' and 'text'. Very briefly, this method is to divide the document image, in block size chosen after a series of tests and then applying the co-occurrence matrix to each block in order to extract five textural parameters which are energy, entropy, the sum entropy, difference entropy and standard deviation. These parameters are then used to classify the image into three regions using the k-means algorithm; the last step of segmentation is obtained by grouping connected pixels. Two performance measurements are performed for both graphics and text zones; we have obtained a classification rate of 98.3% and a Misclassification rate of 1.79%.**

***Keywords*—k-means algorithm, image document, co-occurrence matrix, segmentation, texture.**

I. INTRODUCTION

To switch paper format to an electronic format, we use analysis system and document recognition. Several segmentation techniques documents were published in the literature [12] [22] [25] [26] [7]. These methods can be classified into three approaches namely, bottom-up approaches [14][17] [8], top-down approaches [17] [21] and hybrid approaches [23] [10].

The process using bottom-up techniques are based on the analysis of connected components, these techniques start from pixel level, pixels are then merged into larger components such as homogenous square blocks; connected blocks that have the same characteristics are then merged to form homogeneous regions. The main operators used in this type of approach are thresholding [18], mathematical morphology [19] [1] [20] and projection [10] [9].

When compared to top-down techniques, bottom-up techniques are more efficient when it comes to handling complex layout documents, but have the disadvantage of having a high processing time.

Top-down techniques, proceed by starting from the whole image and split it recursively into different zones until regions in the zone share the same features, like XY cut algorithm that uses the methods of projection profiles [15] and the RLSA algorithm [2] which is based on morphological operations of image processing. Top-down techniques are efficient for good layout structured documents but often fail in complex layout. There are also hybrid techniques that mix the two previously mentioned techniques. Segmentation using texture analysis falls under the latter category.

The methods used separately down and bottom techniques only give good results when the parsed document contains only text, hence the idea of using them together to develop methods called hybrid(mixed). Among these methods we can mention that based texture analysis introduced by Baird [4] that classifies the various components of a document according to the textural characteristics of each zone and method developed by Esposito [3] which consists in applying the smoothing algorithm RLSA with a bottom to classify the blocks according to their content using a decision tree. Other mixed methods exist and most are based on the principle of division and fusion [5] [11] [16] [27]. All these methods can identify three classes namely the 'text', 'background' and 'graphics'. Applications of document recognition cover several areas such as the recognition of codes, street addresses, checks, forms, document archiving and medical bills.

In the next sections we present our method based on co- occurrence matrix for the recognition of the structure of printed and complex documents which presented, “text”, “graphics” and “background”.

II. DESCRIPTION OF METHOD

The technique that we developed to segment a document into three classes (text, graphics, background) is based on the characterization of different regions by textural parameters namely, energy, entropy, sum entropy, difference entropy and standard deviation. These parameters chosen after testing and combination of different Haralick parameters were then used to segment the image

into three regions using the k-means algorithm. The different steps of this technique are summarized in Fig. 1.

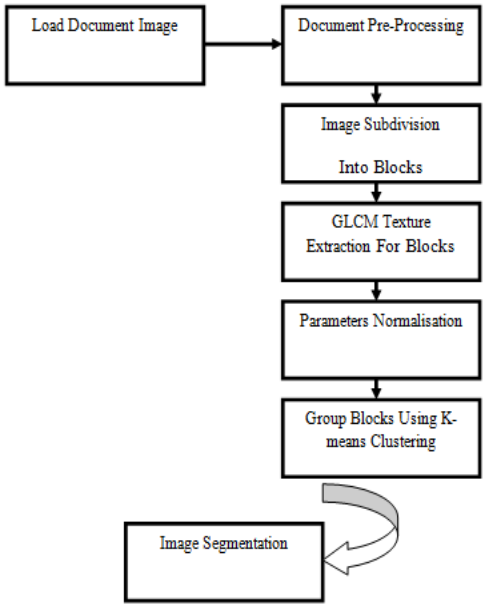

Fig. 1. Flowchart of the system.

A. Image subdivision

The document image is subdivided into blocks, block size is predefined and changes correspond to the size of the image document. Each block becomes the smallest unit for further processing. For an image Img is defined as

$$Img = \{P_{ij}, 0 \leq i < H, 0 \leq j < W\} \quad (1)$$

Where P_{ij} is the pixel of position i, j; H and W are respectively the height and width of the image.
The subdivision of Img into blocks can be expressed as:

$$Img = \left\{b_{IJ}, 0 \leq I < \frac{H}{h}, 0 \leq J < \frac{W}{w}\right\} \quad (2)$$

Where b_{IJ} represents the block of the I_{th} row and J_{th} column of h and w are respectively the height and width of the blocks. A block b_{IJ} is defined as

$$b_{IJ} = \begin{Bmatrix} P_{ij}, h \times I \leq i < h \times (I+1), \\ W \times J \leq j < W \times (J+1) \end{Bmatrix} \quad (3)$$

B. Extraction of textural parameters

The co-occurrence matrix is used to estimate the properties of images related to second-order statistical. This approach is most commonly used to extract texture features [6]. For a translation t, the co-occurrence matrix CM_t of a region R is defined for all pairs of gray level (i, j) as:

$$MC_t = card.\{(s, s+t) R^2 \backslash I(s) = i, I(s+t) = j\} \quad (4)$$

$MC_t(i,j)$ is the number of pairs of sites $(s, s+t)$ of a region, separated by the translation vector t and such that s has the gray level i and $s+t$ has gray level j. For an image Img quantified on Ng gray level, the size of the matrix MC_t is $N_g \times N_g$. $P_{\theta,d}$ is the probability to pass from gray level i to a gray level j of a pitch d(distance between two pixels) and an orientation θ to the horizontal.

$$P_{\theta,d}(i,j) = \frac{CM_t(i,j)}{N} \quad (5)$$

And

$$N = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} CM_t(i,j) \quad (6)$$

For each block of image, five texture parameters are calculated in four directions {0 °, 90 °, 180 °, 270 °}, these parameters are Energy (ENR), Entropy (ENT), Sum Entropy (SEN), Difference Entropy (DEN) and Standard Deviation (STD). Their mathematical definition is given by equations (7-11).

$$ENR = \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} {P_d}^2 (i,j) \quad (7)$$

$$ENT = -\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} P_d(i,j) log_2 P_d(i,j) \quad (8)$$

$$SEN = -\sum_{k=0}^{2n-2} P_{x+y}(k) log_2 P_{x+y}(k) \quad (9)$$

$$DEN = -\sum_{k=0}^{n-1} P_{x-y}(k) log_2 P_{x-y}(k) \quad (10)$$

$$STD = \sqrt{\frac{\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} (P_d(i,j) - \mu)^2}{n \times n}} \quad (11)$$

Where

$$\mu = \frac{\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} P_{d(i,j)}}{n \times n} \quad (12)$$

$$P_{x+y}(k) = \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} P_{d(i,j)} \quad (13)$$

For $i + j = k, k = 0,1 \ldots .2n - 2$

$$P_{x-y}(k) = \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} P_{d(i,j)} \quad (14)$$

For $|i - j| = k, k = 0,1 \ldots . n - 1$

C. Parameters normalization

The scales of individual features can differ drastically. This disparity can be due to the fact that each feature is computed using a formula that can produce various ranges of values. Another problem is that, features may have the same approximate scale, but the distribution of their values has different means and standard deviation. In this work we use statistical normalization (standardization) [24], that independently transforms each feature in such a way that each transformed feature distribution has means equal to 0 and variance equal to 1. A further normalization is performed to enable all the features to have the same range of values that will result in an equal contribution of weight for the similar measure when classifying blocks.

Let P be the number of features and m the size of the distribution, features matrix Z is defined as:

$$Z = \begin{bmatrix} Z_{11} & . & . & Z_{1p} \\ . & . & & . \\ . & & . & . \\ Z_{m1} & . & . & Z_{mp} \end{bmatrix} \quad (15)$$

Where Z_{ij} is the j^{th} feature of the i^{th} candidate for $i = 1,2, \ldots m$ and $j = 1,2, \ldots p$.

The corresponding standardized value is $Z_{ij}{}'$ and is defined in equation (16) as

$$Z_{ij}{}' = \frac{(Z_{ij} - \overline{Z_j})}{\sigma_j} \quad (16)$$

Where $\overline{Z_j}$ is the mean defined in equation (17) and σ_j the standard deviation defined in equation (18).

$$\overline{Z_j} = \frac{1}{m}\sum_{i=1}^{m} Z_{ij} \quad (17)$$

$$\sigma_j = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(Z_{ij} - \overline{Z_j})^2} \quad (18)$$

D. Block classification by the k-means algorithm

To implement the algorithm k-means [13], we set the number of regions to three (k = 3), these regions represent the text, graphics and background. To assign a block to a region, each block of the image is compared to the average value of each class calculated previously.This comparison is performed by minimizing the Euclidean distance between vectors of parameters considered of block and those of the class centers. At each iteration, the algorithm recalculates the center of the classes. This process is repeated until the value of cluster centers does not change.

III. EXPERIMENTAL RESULTS

We used different sizes of test images: $1\,074 \times 820, 1074 \times 768, 1165 \times 850 \ldots$ Test results that led to the choice of size block are given in Table 1. Two performance measurements are performed for both graphics and text zones: Extraction Rate ER and Misclassification Rate MR are defined in equations (19 and 20).

$$ER = \frac{Number\ of\ blocks\ correctly\ extrated}{Number\ of\ Expected\ Correct\ Blocks} \times 100 \quad (19)$$

$$MR = \frac{Number\ of\ misclassified\ blocks}{Number\ of\ Expected\ Correct\ Blocks} \times 100 \quad (20)$$

So the segmented image is divided into M blocks and each block is identified as non-text or text. Then the original image is in turn divided as the same way. Thus we can compare each block of the segmented image to its equivalent in the original image, if the text block observed in the segmented image corresponds to a text block in the original image, so the block is correctly classified but if the text block does not correspond, so the block is misclassified.

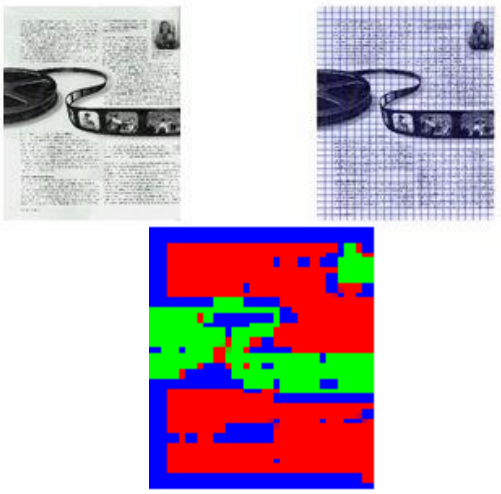

Fig. 2. Division into block and segmentation of image 1.

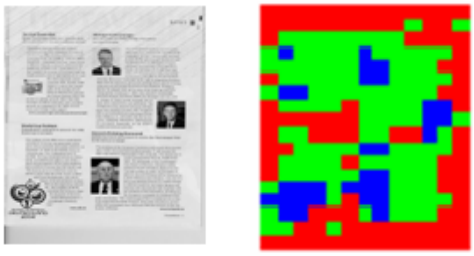

Fig. 3. Segmentation of image 2.

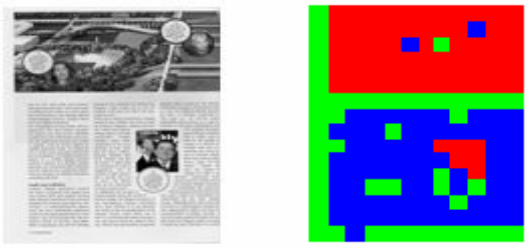

Fig. 4. Segmentation of image 3.

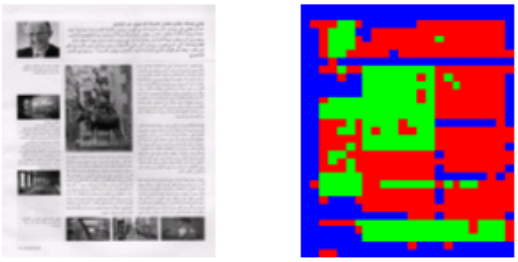

Fig. 5. Segmentation of image 4.

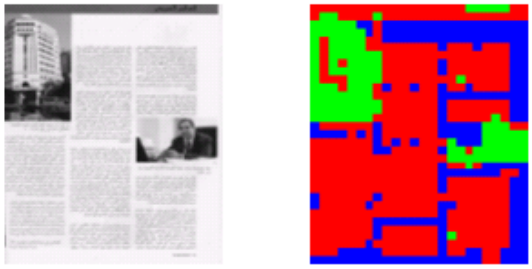

Fig. 6. Segmentation of image 5.

TABLE I.

MISCLASSIFICATION RATE AND COMPUTATIONAL TIME FROM DIFFERENT SIZE BLOCK OF IMAGES.

	M=16		M=32		M=64	
	Time calculation	Relative error	Time calculation	Relative error	Time calculation	Relative error
Img1 768*1074	272.45s	25.35%	91.13s	1.79%	40.05s	12.25%
Img2 786*1074	272.61S	25.25%	91.28s	3.32%	48s	18.68%
Img3 850*1165	344.57s	29.25%	112.12s	1.85%	55.54s	24.00%
Img4 820*1074	290.61s	27.36%	96.17s	3.02%	50.76s	26.18%
Img5 850*1165	322.49s	30.13%	94.33s	2.20%	55.52s	14.53%
Img6 816*1074	290.60s	25.21%	96.10s	2.28%	50.60s	18.72%
Img7 802*1067	281.85s	31.15%	96.40s	2.99%	48.34s	25.57%

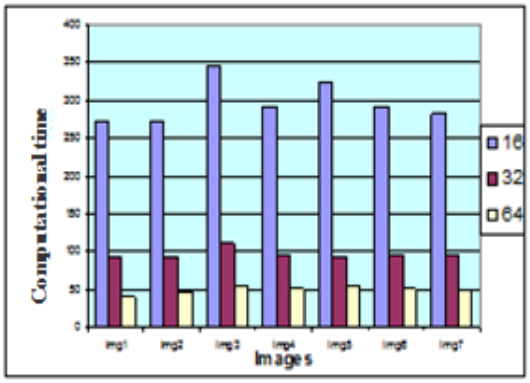

Fig. 7. Comparison of computational time from different size block.

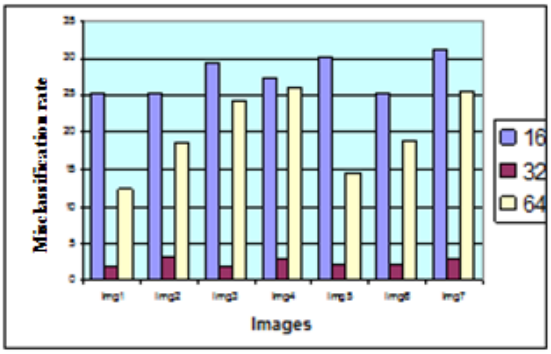

Fig. 8. Comparison of misclassification rate from different size block.

The method developed has been successful from three types of documents processed: documents with non-uniform background and structured with complex layouts (Fig.2), documents containing a text included in the graphics (Fig.4, Fig.6) and documents containing a graphics included in the text(Fig.5,Fig.3).

The result of Fig.2 shows that the three classes have been identified. In this image of document, the green color represents the class 'graphic', the red represents the 'text' and 'blue' represents the 'background. Despite the non-uniform background and complex provisions of the page, the textural parameters of co-occurrence matrix were used to characterize the different textures present in the image. We have obtained a classification rate of 98.21% and misclassification rate of 1.79% from block size of 32*32.

Fig.4 and Fig.6 also show that the three classes of the image document: graphics included in the text,background and text were correctly segmented. Indeed, we have reached for the two images documents a misclassification rates 3.02% and 2.28% respectively.

Fig.3 and Fig.5 also show that the three classes have been identified and we have obtained from the two images documents image a misclassification rates 1.85% and 2.20% respectively.

Test results (Table I.) show that the block size obtained the best result corresponds to 32*32 with a reduced computational time compared to that required using co-occurrence matrix without division block. All results show that our technique developed is well suited to documents with complex structure and significantly improves the classification rate compared to conventional methods based on Gabor filters, the Fourier transform or autocorrelation.

IV. CONCLUSION

The method thus described allowed us to properly separate the different regions of the image document. The tests performed showed the importance of the division block in the image and reduces the computation time especially when using co-occurrence matrix that has very time consuming. We found that the choice of the block size is very important. In fact, the block size should be neither too small as to not contain sufficient information to classify it, nor too great not to include more classes in the same block. After the tests, the best results were obtained for block sizes of 32*32 with misclassification rate of 1.79%. Moreover, the use of image texture as information allowed a good discrimination of the three classes of document image such as text, graphics and the background. As perspective, after segmenting the document, it would be interesting to find a method to extract the different regions of the image document using learning with neural networks.

REFERENCES

[1] N. Amamoto, S. Torigoe, and Y. Hirogaki, (1993) .Block Segmentation and Text Area Extraction of Vertically/Horizontally Written Document. Proc. Second Int'l Conf. Document Analysis and Recognition, p. 739-742, Tsukuba, Japan.

[2] A. Antonacopoulos. Page segmentation using the description of the background. Computer Vision and Image Under standing, 70(3):350–369, 1998. (pages 25, 29 et 39).

[3] A. Azokly. "Une approche générique pour la reconnaissance de la structure physique de documents composites". PhD thesis, IIUF-University of Fribourg, 1995.

[4] H. S. Baird, S. E. Jones and S. J. Fortune. "Image Segmentation by Shape-Directed Covers". Proceedings of the 10th International Conference on Pattern Recognition, Atlantic City, NJ, June 1990, pp. 820-825.

[5] K. Hadjar, O. Hitz and R. Ingold. "Newspaper Page Decomposition using a Split and Merge Approach". Proceedings of the 6th International Conference on Document Analysis and Recognition, Seattle, USA, September 2001, pp. 1186-1189.

[6] R. M. Haralick, K. Shanmugam and I. Dinstein. "Textural features for image classification". IEEE Trans-actions on

Systems, Man and Cybernetics, vol. SMC-3, no. 6, pp. 610–621, November 1973.

[7] A. Jain and S. Bhattacharjee. "Text segmentation using Gabor filters for automatic document processing». Machine Vision and Applications, vol. 5, no.3,

[8] A.K. Jain, Bin Yu (1998) .Document Representation and Its Application to Page Decomposition. IEEE transactions on Pattern Analysis and Machine Intelligence, p. 294-308.

[9] M. Krishnamoorthy and G. Nagy and S. Seth and M. Viswanathan, (1993) .Syntactic Segmentation and Labeling of Digitized Pages from Technical Journals., IEEE Computer Vision, Graphics and image processing, vol. 47, p. 327-352.

[10] Kyong-Ho Lee, Yoon-Chul Choy, and Sung-Bae Cho (Nov. 2000) "Geometric Structure Analysis of Document Images: A Knowledge-Based Approach " IEEE Trans. PAMI, Vol. 22, no. 11, p. 1224-1240.

[11] J. Liu, Y. Tang, Q. He and C. Suen. "Adaptive Document Segmentation and Geometric Relation Labelling: Algorithms and Experimental Results". Proceedings of the 13th International Conference on Pattern Recognition, Vienna, Austria, 1996, pp. 763-767.

[12] Z. Lu. "Detection of Text Regions from Digital Engineering Drawings". IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 4, pp.431–439, 1998.

[13] J. Mac Queen. "Some methods for classification and analysis of multivariate observations". In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. 1967.

[14] Said, J. N. (1997) .Automatic processing of Documents and Bank Cheques. Ph. D Thesis, Concordia University, 452 pages.

[15] G. Nagy and S. Seth. "Hierarchical representation of optically scanned documents". Proceedings of ICPR, 1984, pp. 347-349.

[16] T. Pavlidis and J. Yhou. "Page Segmentation and Classification". CVGIP Vol 54, No. 6, 1992, pp. 482-469.

[17] Ingold, R. (1989) .Structure de documents et lecture optique : une nouvelle approche. Presse Polytechnique Romande, 130 pages.

[18] P.K. Sahoo, S. Soltani, A. K. C. Wong and Y. C. Chen (1980) .A survey of thresholding techniques., Comput. Vision Graph. Image. Press. 41, p. 233-260

[19] J. Serra (1982) .Image analysis and mathematical morphology, Vol. 1, Academic Press, New York.

[20] J. Serra (1988) .Image analysis and mathematical morphology, Vol. 2, Academic Press, New York.

[21] Y.Y. Tang, C.D. Yan, M. Cheriet, and C.Y. Suen (1997) .Automatic analysis and understanding of documents. Handbook of Pattern Recognition and Computer Vision, p. 625-654

[22] K. Tombre, S. Tabbone, L. P´elissier, B. Lamiroy and P.Dosch. "Text/Graphics Separation Revisited". In DAS '02: Proceedings of the 5th International Work-shop on Document Analysis Systems V, pp. 200–211.Springer-Verlag, London, UK, 2002.

[23] Souad Tayeb-bey (1998) .Analyse et conversion de documents : du pixel au langage HTML. Thèse de Doctorat, INSA de Lyon, France, 175 pages.

[24] S. Teeuwsen. "Feature selection for small-signal stability assessment". In Proceedings of the Dresdner Kreis 2002. Werningerode, Germany, March 2002.

[25] Q. Yuan and C. Tan. "Text extraction from gray scale document images using edge information". In Proceedings of the ICDAR 2001, pp. 302–306. 2001.

[26] Y. Zheng, H. Li and D. Doermann. "Machine printed text and handwriting identification in noisy document images". Tech. rep., LAMP Lab, University of Maryland, College Park, 2002.

[27] Y. Zhong, K. Karu, and A. K. Jain. Locating text in complex color images. Pattern Recognition, 28(10):1523–1535, 1995. (page 33.)