Text/Image separation in document images based on statistical analysis of texture and Morphological operations

Fattah Zirari, Driss Mammass

IRF-SIC Laboratory Ibn Zohr University Agadir, Morrocco

[zirari\_fattah@yahoo.fr](mailto:zirari_fattah@yahoo.fr) , [mammass@univ-ibnzohr.ac.ma](mailto:mammass@univ-ibnzohr.ac.ma)

Abdellatif Ennaji LITIS Laboratory University of Rouen

Rouen, Farnce

[Abdel.ennaji@univ-rouen.fr](mailto:Abdel.ennaji@univ-rouen.fr)

***Abstract*—The separation of text / image is a major step in the processing of documents images. It consists of separating the document into two classes: text and image. In this context, it is important to implement approaches that can handle such documents.**

**This paper presents a new method of separating text / image into a document image. The method developed is based on statistical analysis of texture and morphological operations. This is, initially, to calculate the energy for each pixel, using a method to extract features such as co-occurrence matrix and we have applied the morphological operations. we have tested this method on a database of 150 newspapers. The experimental results show satisfactory values and are very promising.**

***Keywords— segmentation ; texture analysis ; co-occurrence matrix ; morphological operations; document image.***

1. Introduction

Segmentation is a basic step of the image processing that precedes any other identification or classification. This step depends on the type of image that differs from both by the acquisition system and the process of image formation. In the case of document images, this means identifying and locating the two areas of interest: text and image. Two approaches are possible to determine this structure, a bottom-up or a top-down approach.

The first is to extract the primitive documents, and then iteratively combine these primitives in homogeneous region. . The main operators used in this type of approach are thresholding [12], mathematical morphology [6], [11], [7] and

projection [9] [10].

The second seeks to recursively divide the image into homogeneous sub-regions. Among the techniques used in this type of approach, we can cite the XY cut algorithm that uses the methods of projection profiles [4] and the RLSA algorithm

[1] which is based on morphological operations of image processing.

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.

Finally a third approach consist of combining the two approaches. Among the techniques used in this type of approach we can mention those based on texture analysis [2] which allows to classify the different components of a document under the textural characteristics of each zone and one developed by Esposito [3] of applying the smoothing algorithm RLSA with a bottom-up approach to classify the blocks according to their content using a decision tree. Other techniques are mixed and most are based on the principle of split and merge [8] [5] [14] [15].

In this paper we present the successive steps for segmenting a document image into two class : text and image, we focus specifically on statistical analysis of texture by co-occurrence matrix. The remainder of the paper is organized as follows: Section (2), we describe steps used by our approach. Next, we present experimental results in Section (3). Finally, the conclusions are presented in Section (4).

1. Method description

The method suggested realizes the separation text/image in a document multi image. This method which we developed to segment a document in two areas of interests: the text and the image.

The different steps of this technique are summarized by the flowchart in the following figure:

978-1-61284-732-0/11/$26.00 ©2010 IEEE

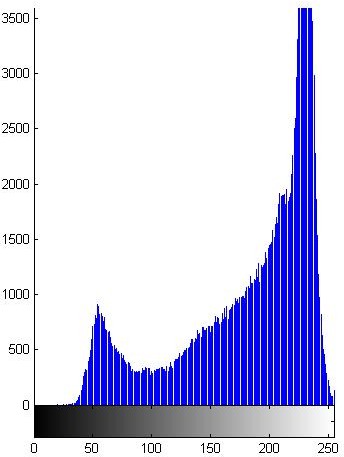
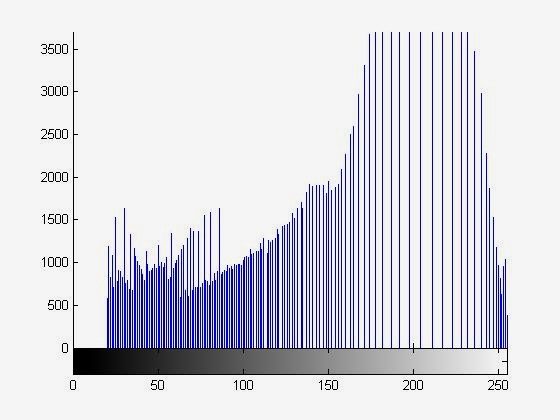
 

Figure 1. Flowchart of the method



Extraction the energy in each pixel by co- occurrence matrix

Morphological operations:

Text

oponing and closing

Image

Binarization by Otsu

Document pre-Pre-processing: Histogram Equalization

Load document image

1. *Pre-processing: histogramm equalization*

Pre-processing consists of those operations that prepare data for subsequent analysis that attempts to correct or compensate errors. After pre-processing is complete, the analyst may use feature extraction to reduce the dimensionality of the data.

The image enhancement is performed by a set of techniques that aim to improve the appearance of the image into a form more accessible to analysis. All the techniques developed in this direction must preserve the essential properties of shapes, which could lead otherwise, serious errors of analysis and later recognition.

c) d)

Figure 2. a) Original document, b) Document after enhancement, c) Histogram of original image, d) Histogram after equalization

1. *Computing the energy in each pixel by co-occurrence matrix*

Co-occurrence matrix, one of the most known texture analysis methods, estimates image properties related to second- order statistics. Each entry *(i,j)* in co-occurrence matrix corresponds to the number of occurrences of the pair of gray levels *i* and *j* which are a distance d apart in original image. The main caracteristic of co-occurrence matrix is to look at pairs of pixels that are, by definition, separated by a distance shown later in this paper by the vector *d*.

A discrete image is regarded as a function of two variables on a discrete domain D, of dimension *M x N*. The function *f* takes its values in the discrete set of *K* elements *E*, *E = {0, 1,. .*

*. . K - 1}* and is written:

In our case, the preprocessing consists of increasing the contrast of images by histogram equalization.

*f* : *D*  *f* (*x*, *y*)  *E*

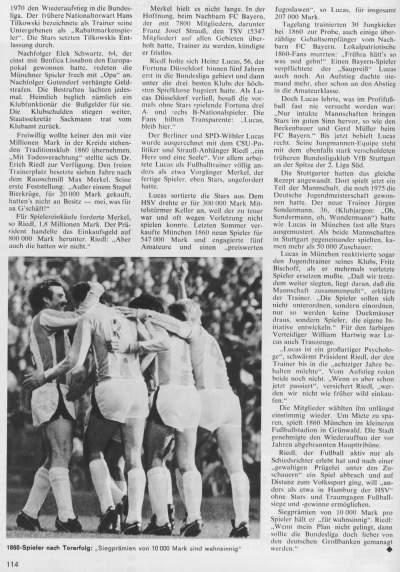
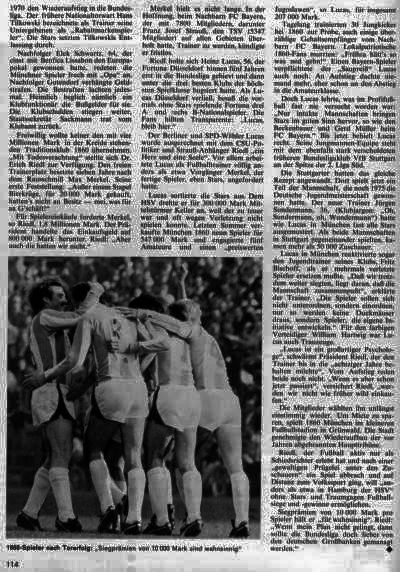
**(1)**

The contrast enhancement using histogram equalization is a specific operation based on an analysis of the histogram of gray levels of the source image, and automatically corrects the distribution of gray levels to use the entire dynamic levels of gray. The goal of histogram equalization is to reduce the gray level histogram to a histogram indicating a flat equiprobability different level of gray.

The values of *f (x, y)* are the intensities of image pixels. The co-occurrence matrix is a parameterized function of *E x E* into

*N*. we denote *i* and *j* the two variables co-occurrence matrix. The parameters of the matrix are the image f and the displacement vector expressed by *[dx, dy]* in Cartesian coordinates by *[r = | d |, θ = arctan (dy / dx)]* in polar coordinates.

The co-occurrence matrix is thus defined by:

*c* : *E*  *E*  *N*

( *i* , *j* ) 

*c* ( *i* , *j* , *f* , *d* ) 

*Aij*

**(2)**

With

*Aij*

the cardinal of the set

*Aij* defined by:

*x*, *y*, *x*, *y* *x*, *y*  *D*  *t**d* , *x*, *y*  *D*  *td* 

*Aij*  

**(3)**

 *et* *x*, *y*  *td* *x*, *y*  *et f* *x*, *y*   *i et f* *x*, *y*  *j*

In this expression *td* is the vector of translation *d*. The cardinal

a) b)

*Aij*

is equal to the number of pairs of pixels separated by the

displacement and such that the pixel intensities are the first *i* and *j* respectively. With this formulation a couple of pixels of intensities *i* and *j* is included in the calculation *c* *i*, *j*, *f* , *d*  and

not in that *c*  *j*, *i*, *f* , *d*  . The co-occurrence matrix is not symmetric. The computed matrix will thus have information depending on the direction of vector *d*. A second set, denoted

information contained in the co-occurrence matrix and allow better discrimination of textures. The most relevant feature that is widely used in literature is:

Energy, also called Angular Second Moment and Uniformity, is a measure of textural uniformity of an image. Energy reaches its highest value when gray level distribution has either a constant or a periodic form. A homogenous image

*A i**j*

is constructed by replacing the translation vector by a

contains very few dominant gray tone transitions, and therefore

translation of vector **-** *d*. Symmetrical co-occurrence matrix is then expressed by:

the P matrix for this image will have fewer entries of larger magnitude resulting in large value for energy feature. In contrast, if the P matrix contains a large number of small entries, the energy feature will have smaller value.

*C s* : *E*  *E*  *N*



*i* , *j*   *C s ij* :

*A ij* 

*A i**j*

**(4)**

*E*    *P*

*i j*

*i* ,

*j* 2

**(7)**

The expression (3) indicates that pixels in the set

*Aij*

1. *Morphological operations : oponing and closing*

belongs to a domain definition *D*  *td* *D*  included in *D*. The reduction of the computational domain is the result of the use of the translation *t d* . To allow comparative analysis between

images, the coefficients *Csij* are standardized. Each coefficient is divided by the sum of coefficients of co-occurrence matrix. This sum is noted *A* and corresponds to the number of pairs of pixels taken into account in calculation, this number is also, by the discretization of domains, the domain calculation size or their cardinal, is:

Morphology is a broad set of image processing operations that process images based on shapes. Morphological operations apply a structuring element to an input image, creating an output image of the same size. In a morphological operation, the value of each pixel in the output image is based on a comparison of the corresponding pixel in the input image with its neighbors. By choosing the size and shape of the neighborhood, you can construct a morphological operation that is sensitive to specific shapes in the input image.

The most basic morphological operations are dilation and erosion. Dilation adds pixels to the boundaries of objects in an

*A*  *D*  *t d* ( *D* ) 

*D*  *t*  *d* ( *D* )

image, while erosion removes pixels on object boundaries. The

 2  *M*

* *d x*

 *N*  *d y* 

**(5)**

number of pixels added or removed from the objects in an image depends on the size and shape of the structuring element used to process the image.

The full expression of the co-occurrence matrix, symmetric

and normalized, is the following:

Opening and closing are compound operators that are, respectively, morphological erosion followed by dilation, and

*C* : *E*  *E*  *R*

*A ij* 

*A i**j*

**(6)**

dilation followed by erosion.

*i* ,

*j*  

*C ij*  *A*

III.

RESULTATS

The co-occurrence matrix show the relation which exist between pixels on a local aspects (levels of gray) and a spatial aspect (displacement). However, this is true only if a large number of matrix are calculated. If we limit ourselves to the calculation of some matrix, by considering only one direction and some distance (and vice versa), most of information well be lost (one is likely to omit orientations if there was not enough directions, or to restrict spatially with limited distances). This shows that, in addition to the importance of the choice of displacement, it remains to calculate a rather significant number of matrix.

The co-occurrence matrix contain a considerable mass of information and are therefore difficult to handle, so fourteen features (defined by Haralick) [13] which correspond to descriptive character textures can be calculated from these matrix. These features, although correlated, reduce the

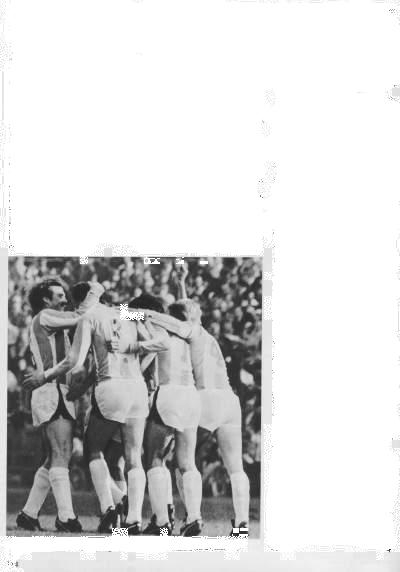
In this section we present the results obtained, first, with the statistical analysis of texture. Thereafter, we show the interest of the use of the morphological operations: opening and closing to refine the result of the text/image separation in document image.

At first sight, we note that we have a great difference between the results; the text/image separation resulting from combining statistical analysis of texture-Morphological operations is visually more satisfactory than those of statistical analysis of texture since it is less noisy. For the statistical analysis of texture-Morphological operations method a total accuracy of 98.90% was achieved, slightly higher than the accuracy of 92.61% derived by the statistical analysis of texture (detailed in the following table).

TABLE I. TOTAL ACCURACY ACHIEVED

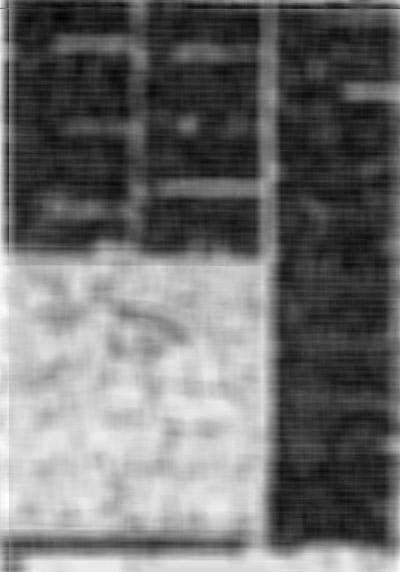
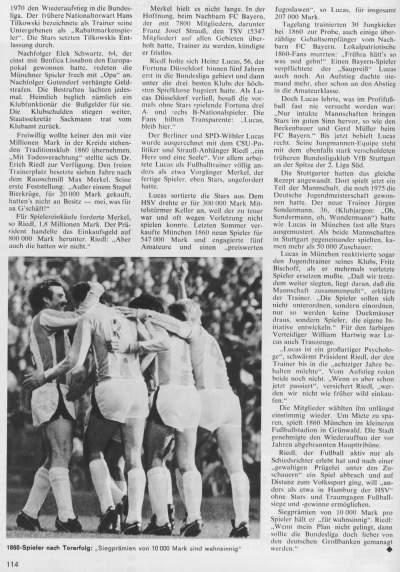
Method statistical analysis of texture

statistical analysis of texture-Morphological operations



ACCURACY 92.61% 98.90%

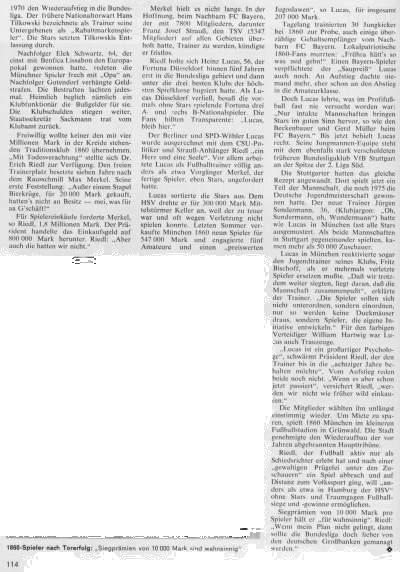
ACHIEVED



a) b)



c) d)



e) f)

g) h)

Figure 3. a) Original document, b) Computing the energy, c) Binary document d) Binary document after morphological operations, e) Text extraction, f) Text extraction after morphological operations, g) Images extraction, h) Images extraction after morphological operations

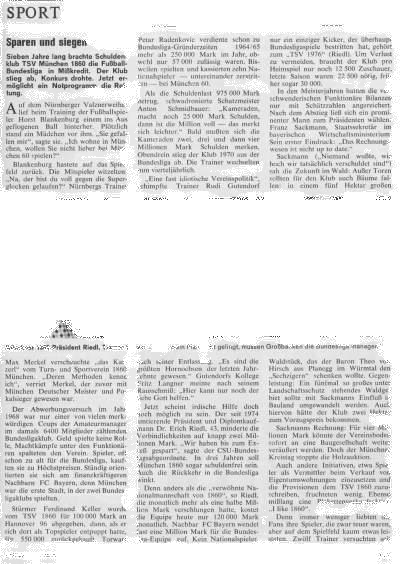
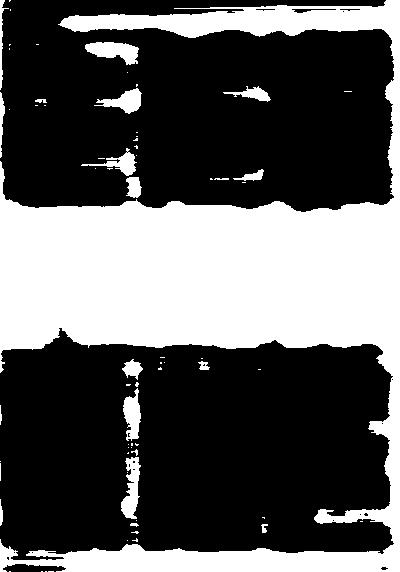
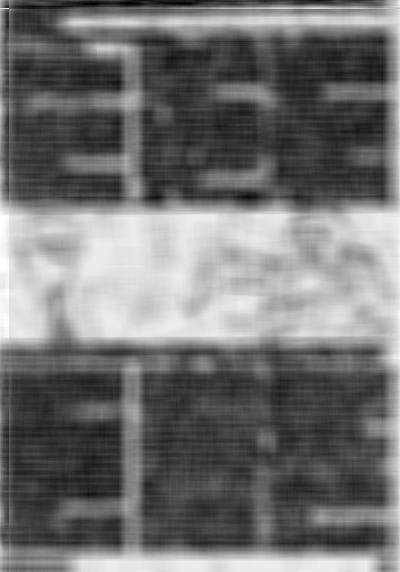
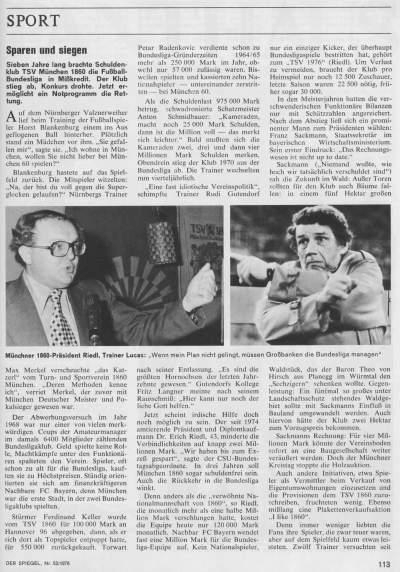


Figure 4. result before morphological operations



Figure 5. result after morphological operations

IV. Conclusion

We segmented document images by applying an approach based on statistical analysis of texture and Morphological operations. The application of this methodology gives more satisfactory results. Indeed, the method as described has allowed us to separate adequately the different regions of the image document: text and image.

Acknowledgment

We would like to acknowledge the financial support of our project by to “action intégrée Maroc-francaise” n° MA/10/233 and the AIDA project, program Euro Mediterranean 3+3 : n° M/09/05.

References

1. A.Antonacopoulos. Page segmentation using the description of the background. Computer Vision and Image Understanding, 70(3):350– 369, 1998. (Cité pages 25,29 et 39.) .
2. A.Azokly. “Une approche générique pour la reconnaissance de la structure physique de documents composites”. PhD thesis, IIUF- University of Fribourg, 1995.
3. F.Esposito, D. Malerba and G. Semeraro. “A Knowledge-Based Approach to the Layout Analysis”. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, Canada, 1995, pp. 466-471.
4. G. Nagy and S. Seth. “Hierarchical representation of optically scanned documents”. Proceedings of ICPR, 1984, pp. 347-349.
5. J.Liu, Y.Tang, Q. He and C.Suen. “Adaptive Document Segmentation and Geometric Relation Labeling: Algorithms and Experimental Results”. Proceedings of the 13th International Conference on Pattern Recognition, Vienna, Austria, 1996, pp. 763-767.
6. J.Serra (1982) .Image analysis and mathematical morphology, Vol. 1, Academic Press, New York.
7. J.Serra (1988) .Image analysis and mathematical morphology, Vol. 2, Academic Press, New York.
8. 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.
9. 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.
10. M.Krishnamoorthy and G. Nagy and S. Seth and M. Viswanathan, (1993) .Syntactic Segmentation and Labeling of Digitized Pages from Technical Journals., IEEEComputer Vision, Graphics and image processing, vol. 47, p. 327-352.
11. N.Amamoto, S.Torigoe, and Y.Hirogaki, (1993) .Block Segmentation and Text Area Extraction of Vertically/Horizontally Wrtitten Document., Proc. Second Int’l Conf. Document Analysis and Recognition, p. 739-742, Tsukuba, Japan.
12. P.K. Sahoo, S. Soltani, A. K. C. Wong and Y. C. Chen (1980) .A survey of thresholding techniques., Comput. Vision Graph. Imag. Press. 41, p. 233-260.
13. R.M. Haralick, K. Shanmugam, and I. Dinstein, “Textural Features for Image Classification”, IEEE Trans. on Systems, Man, and Cybernetics, Vol. SMC-3, No.6, November 1973, pp.610-621.
14. T. Pavlidis and J. Yhou. “Page Segmentation and Classification”. CVGIP Vol 54, No. 6, 1992, pp. 482-469.
15. Y.Zhong, K. Karu, and A. K. Jain. Locating text in complex color images. Pattern Recognition, 28(10) :1523–1535, 1995. (Cité page 33.) .