Page segmentation based on Steerable Pyramid features

Mohamed Benjelil
REGIM – ENIS, B.P. 1173, 3038, Sfax, Tunisia
L3I, Univ. of La Rochelle, France
mohamed.benjlaiel@ieee.org

Rémy Mullot
L3I, University of La Rochelle, Avenue Michel Crépeau, 17042. La Rochelle, France
remy.mullot@univ-lr.fr

Adel M. Alimi
REGIM – ENIS, B.P. 1173, 3038, Sfax, Tunisia
adel.alimi@ieee.org

Abstract

Page segmentation and classification is very important in document layout analysis system before it is presented to an OCR system or for any other subsequent processing steps. In this paper, we propose an accurate and suitably designed system for complex documents segmentation. This system is based on steerable pyramid transform. The features extracted from pyramid sub-bands serve to locate and classify regions into text (either machine printed or handwritten) and non-text (images, graphics, drawings or paintings) in some noise-infected, deformed, multilingual, multi-script document images. These documents contain tabular structures, logos, stamps, handwritten script blocks, photos etc. The encouraging and promising results obtained on 1,000 official complex document images data set are presented in this research paper.

1. Introduction

In the face of the very important mass of information exchanged between different organizations, the need for systems allowing the recognition, the indexation, the information retrieval and the automatic classification of complex multi-lingual and multi-script document images has grown continuously. However, most works of retro-conversion of printed Arabic document images are limited to textual block recognition without handling complex documents such as letters of information, forms, all types of application sheet, etc. In practice, these documents can be noised, skewed, deformed, multi-lingual, multi-script, with irregular textures and may contain several heterogeneous blocks such as annotations, machine print and/or handwritten script, graphics, pictures, logos, photographs, tabular structures. This situation makes it difficult to analyze and recognize document images.

Various schemes of page segmentation have been proposed by researchers. One of the most well known approaches for page segmentation is the one of connected components aggregation [2]. Connect component based algorithms fall in the category of bottom-up segmentation algorithms. Similar components are iteratively grouped together to form progressively higher-level descriptions of the printed regions of the document such as words, lines, and paragraphs [3]. Other approaches use the description of white space to identify homogeneous regions. In most of the other approaches in the literature, researchers make assumptions about the general layout of the document page to be segmented. Some assume that the text or image blocks may only be rectangular [4], while others may assume that sentences in text are all evenly spaced. Others assume that the document belongs to a specific category, such as a newspaper [5] or a technical article [6]. In [7], a text-image-structure-analysis, analogous to a document structure analysis, is needed to enable a text information extraction system to be used for any type of image, including both scanned document images and real scene images.

However, making such assumptions restricts the applicability of the page segmentation scheme to a limited number of document classes. In [8], most of the work makes use of deterministic models. Such models fail in the presence of noise or ambiguity.
In [9], the authors proposed a modification of two existing texture based techniques such as Gabor feature based method & Log polar wavelet signature method with the inclusion of Harris corner detectors for Document images. Even though the previous approaches present some advantages, some of them were evaluated under controlled scenarios. Furthermore, typical classification processes utilizing only the original input image at a fixed feature scale are sensitive to noise and will lead to intra-object classification errors.

In this paper, we propose a new approach to the non-controlled environments document image segmentation which is based on steerable pyramid decomposition. We essentially focus our interest on the segmentation of

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.253

complex document images in text and non-text regions.

The rest of the paper is organized as follows: overview of the steerable pyramid decomposition is presented in section two. In Section three, we present our segmentation strategy. Experimental results and performance analysis are presented in section four, and future works and conclusion are provided in the last section.

2. Steerable pyramid (S.P)

Steerable Pyramid Decomposition [10, 11], is a linear multi-orientation, multi-resolution image decomposition method, by which an image is subdivided into a collection of sub-bands localized at different scales and orientations.

The synoptic diagram for a first level image decomposition using Steerable Pyramid is shown in figure (1-a). Using a high-pass and low-pass filter (*H0, L0*) the input image is initially decomposed into two sub-bands: a high-pass, and a low-pass sub-band, respectively. Further, the low-pass sub-band is decomposed into *K*-oriented band-pass portions *B*0 to *BK*−1, and into a low-pass sub-band *L*1. The decomposition is done recursively by sub-sampling the low-pass sub-band by a factor of 2 along the rows and columns. Each recursive step captures different directional information such as variation of a texture in both intensity and orientation at a given scale. Figure (1-b) shows the image decomposition into 3 levels and 4 orientations (0°, 45°, 90°, and 135°).

The basic functions of the steerable pyramid are directional derivative operators that come in different sizes and orientations. The simplest example of this is the oriented first derivative of Gaussian. Consider the two-dimensional circularly symmetric function G written in Cartesian coordinates x and y:

$$G(x, y) = e^{-(x2+y2)} \quad (1)$$

The scaling and normalization constants have been set to 1 for convenience. The directional derivative operator is steerable as is well-known [11], Let us write the *n*th derivative of a Gaussian in the *x* direction as G_n. Let $(\ldots)^{\theta}$ represent the rotation operator such that for any function $f(x, y)$, $f^{\theta}(x, y)$ is $f(x, y)$ rotated through an angle θ about the origin. The first *x* derivative of a Gaussian $G_1^{0^\circ}$ is

$$G_1^{0^\circ} = \frac{\partial}{\partial x} e^{-(x^2+y^2)} = -2xe^{-(x^2+y^2)} \quad (2)$$

The same function, rotated 90°, is

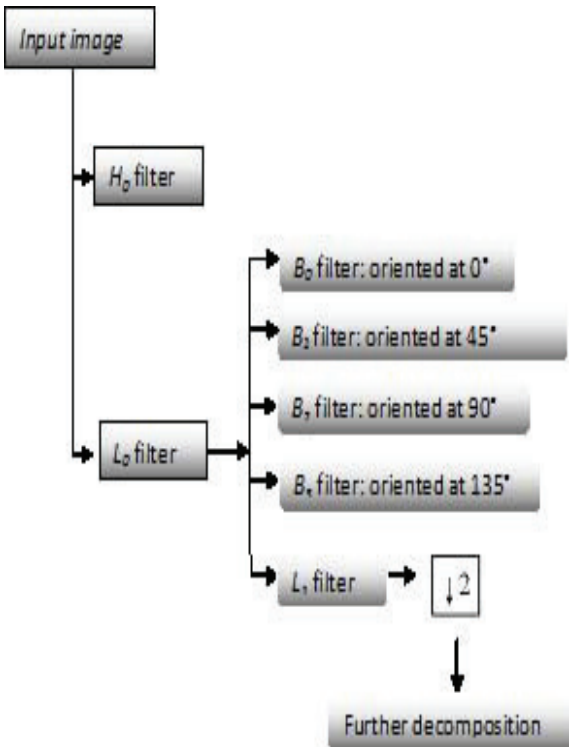
$$G_1^{90^\circ} = \frac{\partial}{\partial y} e^{-(x^2+y^2)} = -2ye^{-(x^2+y^2)} \quad (3)$$

It is straightforward to show that a G_1 filter at an arbitrary orientation θ can be synthesized by taking a linear combination of $G_1^{0^\circ}$ and $G_1^{90^\circ}$.

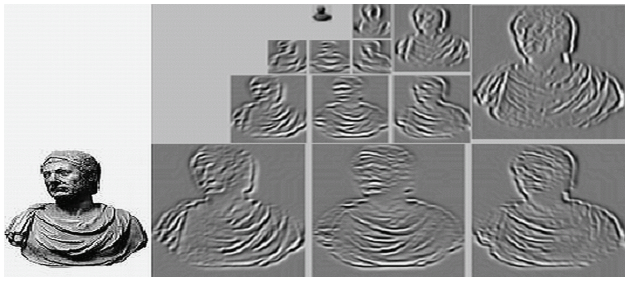
$$G_1^{\theta} = \cos(\theta)\, G_1^{0^\circ} + \sin(\theta)\, G_1^{90^\circ} \quad (4)$$

Since $G_1^{0^\circ}$ and $G_1^{90^\circ}$ span the set of G_1^{θ} , we call them basis filters for G_1^{θ}. The $\cos(\theta)$ and $\sin(\theta)$ terms are the corresponding interpolation functions for those basis filters. Because convolution is linear operation, we can synthesize an image filtered at any arbitrary orientation by taking linear combinations of the images filtered with $G_1^{0^\circ}$ and $G_1^{90^\circ}$. Letting * represent convolution and *I* the input image, for $R_1^{0^\circ} = G_1^{0^\circ} * I$ and $R_1^{90^\circ} = G_1^{90^\circ} * I$, the resulting image is $R_1^{\theta} = \cos(\theta)\, R_1^{0^\circ} + \sin(\theta)\, R_1^{90^\circ}$ (5)

Figure (2) shows the synthesized image.



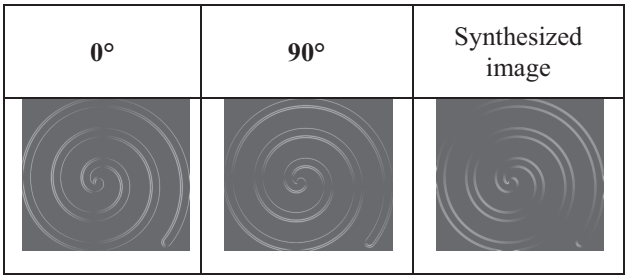
(a)



(b)

Figure 1. a) First level of steerable pyramid decomposition [11] b) Image decomposition into 3 levels and 4 orientations (0°, 45°, 90°, and 135°)

(a)



(b)

Figure 2. a) Input image b) Synthesized image by taking linear combinations of the images filtered with $G_1^{0^\circ}$ and $G_1^{90^\circ}$ [1]

We tested the S.P on 300 images of text blocks and 300 images of non-text objects (images of natural scene). For each image sub-bands, we calculated the mean and standard deviation. Each *marker* in the scatter plot, Figure (3), represents an observation, and its position shows the values of mean and standard deviation for that observation. This figure shows how the statistical measurements differ between text and non-text objects. This means that the S.P features have a distinguishing capability between text and non-text objects. Since the complex document images are basically constituted by such types of objects, we decided to use the S.P features in the classification part of our segment system.

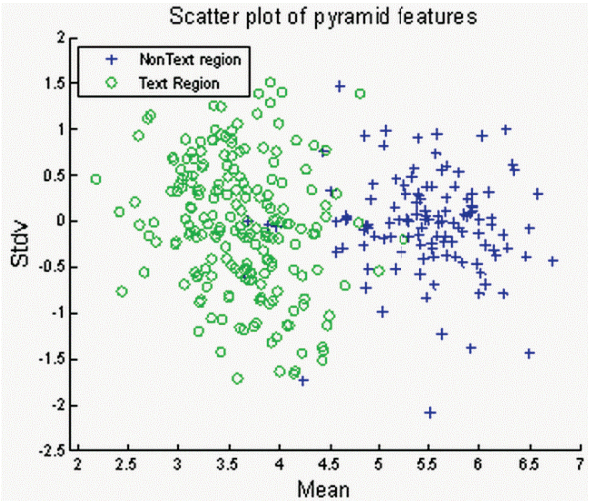


Figure 3. Pyramid features scatter plot (o) text blocks (+) non-text blocks

3. Segmentation strategy

We consider text and non text as regions with different textures. Since the distinguishing characteristics of text are frequency information, orientation, approximately with the same size and line thickness, located at a regular distance from each other, we can use them to characterize machine-printed text regions. Handwritten script and annotations detection is more difficult to implement than machine-printed text due to the diversified human handwriting styles and customs. To overcome this problem our system must be robust to transformations, with efficient feature extraction and with a representative training data set.

In order to satisfy the two first conditions, we used the steerable pyramid decomposition due to its invariance to translation, rotation, and scaling [11]. To satisfy the third condition, we used a representative training data set containing, in addition to machine print text, handwritten script with different sizes, orientations and writing styles. Figure (4) shows the synoptic diagram of proposed system.

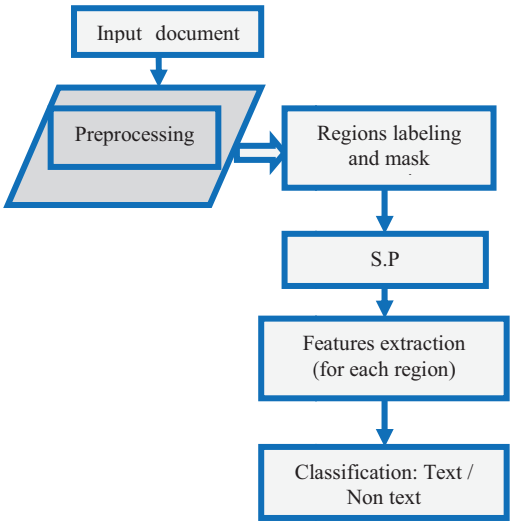


Figure 4. Synoptic diagram of proposed segmentation system

3.1. Preprocessing

Some preprocessing operations are required before the ensuing document segmentation. Firstly, document content must be segmented from the background and we adopt Otsu's [16] global thresholding technique for document binarization. Second, we used a classical filtering technique such as morphological operators that remove isolated small objects without removing texts diacritic points.

3.2. Regions labeling, mask generation and application

Since text shows spatial cohesion, characters appear in clusters at a regular distance aligned to a virtual line. By merging characters inside each cluster by image dilatation with suitable horizontal and vertical structuring elements we can create a document mask figure (5b).
This mask will serve to extract corresponding regions from original image. We applied the steerable decomposition on each region. Each region will be classified as text or non-text based on feature values obtained from pyramid sub-bands. Figure (5c) shows some samples of words and images and their corresponding S.P decomposition. Figure (5d) shows the final result.

(a) (b) (c) (d)

Figure 5. Segmentation process a) Input image b) Image mask c) Samples of words and images and their corresponding S.P decomposition d) Final result with Blue boundaries as text, and Magenta boundaries as non-text

3.3. Features extraction

The document image is decomposed into 3 levels (1, 2 and 3) and 4 orientations (0°, 45°, 90°, and 135°) for sub-bands, using the steerable pyramid transforms. This configuration gives 12 sub-bands (3 levels x 4 orientations). The feature vector equation (6) (with dimension 48 to represent 12 sub-bands) was constructed based on the computed mean μ_{mn}, the standard deviation σ_{mn} of the transformed region image and the energy E_{mn}, the homogeneity H_{mn} calculated from gray-level co-occurrence matrix applied to the same transformed region image. This feature vector is defined as:

$$FV = [\mu_{11}\ \sigma_{11}\ E_{11}\ H_{11}\ \ldots\ \mu_{12}\ \sigma_{12}\ E_{12}\ H_{12}] \quad (6)$$

3.4. Classification

To classify the extracted regions into text or non-text classes, we tested three classification methods. In the first method, we used a non-supervised k-means classifier with two classes based on regions detected in a single document, one for the text and the other for the non-text. We consider the class that includes more items as the text class. This is justified by the idea that in a document, the number of textual regions is greater than the number non-textual ones. In the second method, we used a supervised K nearest neighbor's classifier (KNN) with different values of K (3, 5, and 7). In this framework, starting with 4,664 text block images and 1,258 graphic blocks, we choose 600 text block images and 600 graphic blocks as training data set. Figure (5c) shows some samples of extracted regions and their corresponding S.P transform, Figure (5d) shows the final segmentation result. In the third method, we used a Naïve Bayes classifier with reject Option.

4. Results and performance analysis

4.1. Experimentation Setup

This algorithm has been tested over a corpus of 1,000 images with 3 sets of images and various metrics have been evaluated from tested results.

- Set1 contains 350 Dispatch notes.
- Set2 contains 350 Forms
- Set3 contains 300 newspapers

The S.P parameters used are sp3filter with 3 levels (scales) and 4 orientations [11].

4.2. Performance Analysis

Metrics used to evaluate the performance of the system are Precision rate (P), Recall rate (R) and F-Score. Precision and Recall rates have been computed based on the number of correctly detected regions in an image in order to evaluate the efficiency and robustness of the algorithm and the Metrics are as follows:

$$P = \frac{\text{Correctly detected regions(True positives)}}{\text{Correctly detected regions+False positives}} * 100\% \quad (24)$$

$$R = \frac{\text{Correctly detected regions}}{\text{Correctly detected regions+False negatives}} * 100\% \quad (25)$$

F-score is the harmonic mean of recall and precision rates as represented in equation (26).

$$F - score = 2 * \frac{P*R}{(P+R)} \quad (26)$$

4.3. Results

In Figure (6), we tested our segmentation method for different document images types: original images are on the left and segmented image mask are in the middle. The output of the text / non-text segmentation is on the right. The first example is for a noisy official document image with skewed handwritten annotations. The second document image is for a magazine with non-Manhattan layout.

From Table (1), we can see that the naïve Bayes classifier achieves the highest accuracy with F-score 95.03%. The k-means follows it with F-score 93.44%. The knn classifier with K=5, achieves the lowest accuracy with F-score 88.56%. The segmentation results by type of document with k-means, KNN and Bayes classifier are presented in Table (2), Table (3), and Table (4) respectively.

The overall segmentation rate is about 92.34 %. This is reasonable at this stage and we can afford adding some preprocessing and more improved feature selection.

Table 1. Correct classification rates

Classifier	R	P	F-score
k-means	94.83 %	92.33 %	93.44%
knn	89.26 %	86.79 %	88.56%
Naïve Bayes	96.42%	93.68%	95.03%

Table 2. Correct segmentation rates by type of document with k-means classifier

Documents	Nb	R	P	F-score
Dispatch notes	350	93.80%	94.60%	94.16%
Forms	350	99.20%	93.90%	96.46%
News paper	300	91.50%	88.50%	89.70%

Table 3. Correct segmentation rates by type of document with KNN classifier with K=5

Documents	Nb	R	P	F-score
Dispatch notes	350	88.98%	88.92%	88.51%
Forms	350	93.24%	88.26%	91.04%
News paper	300	85.58%	83.19%	84.31%

Table 4. Correct segmentation rates by type of document with Naïve Bayes

Documents	Nb	R	P	F-score
Dispatch notes	350	97.96%	98.12%	98.03%
Forms	350	99.46%	96.68%	98.05%
News paper	300	96.36%	92.24%	94.25%

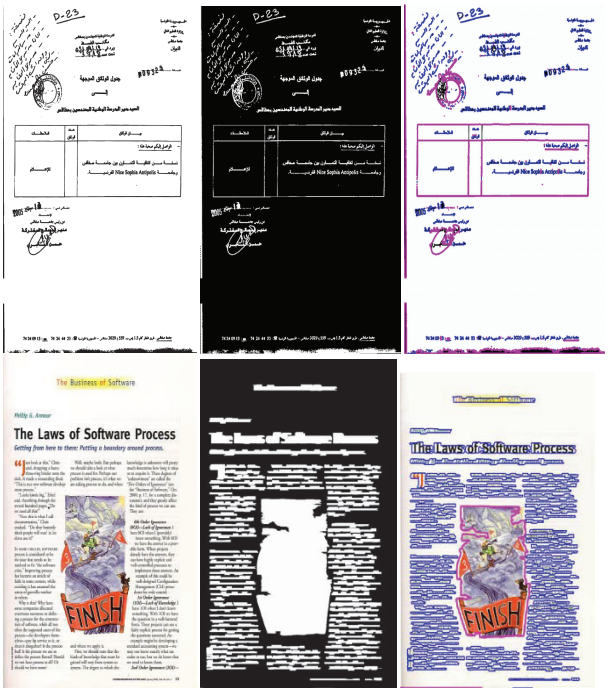


Figure 6. First column: Original image, second column: segmented image mask, third column: image segmented into text and non-text regions with concave hull. The color bar for this figure is blue for text and Magenta for non-text.

4.4. Comparison with other segmentation techniques

In order to evaluate the performance of our system, we compared it with Docstrum [12] and X-Y cut [13] algorithms for page segmentation. The performance evaluation method used is based on counting the number of matches between the entities detected by the algorithm and the entities in the ground truth [15].
A performance metric for detecting each entity can be extracted if we combine the values of the entity's detection rate and recognition accuracy. We can define the following Entity Detection Metric (EDM):

$$EDM = 2 * \frac{DetectionRate*RecognitionAccuracy}{DetectionRate+RecognitionAccuracy} \quad (30)$$

We have randomly selected dozens of document images from MediaTeam Document Database [14], from the Page Segmentation Competition of ICDAR2009 [15]. The dataset is divided into 100 training images and 200 test images.
The evaluation results of the 3 segmentation algorithms are shown in Figure (7) where the EDM values averaged over all images are depicted. In this figure we see that the

proposed method achieved the highest averaged detection rate, recognition accuracy and EDM rate values (86.96%, 83.33% and 85.11% respectively).

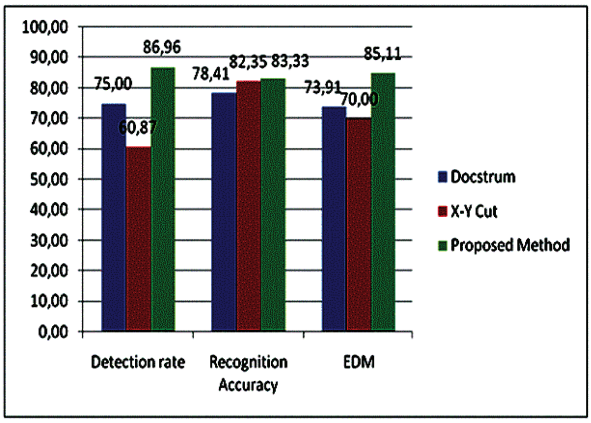


Figure 7. Evaluation results for all entities (EDM values averaged over all images)

5. Conclusion and future work

The work developed in this paper aims at setting up a system of segmentation of complex multilingual multi-script document images. Thus, we began with a study of the existing systems of document images segmentation. Within this framework, we showed a few systems that handled complex multilingual multi-script document images. We presented the proposed system which is based on steerable pyramid. Lastly, we exposed the results obtained on a data set of 1,000 documents. The rate of correct segmentation obtained is about 93.44 %. Additional efforts need to be made on multi-oriented annotations detection and the separation of graphic linked to text

References

[1] W. T. Freeman, E. H. Adelson The design and use of steerable filters. IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 13, no. 9, pp. 891-906, September, 1991.

[2] Fletcher L., Kasturi R., `A Robust Algorithm for Text String Separation From Mixed Text/Graphics Images', IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 10, pp.. 910-918, 1988

[3] Tan C. L., Yuan B., Huang W., Zhang Z., `Text/Graphics Separation using Pyramid Operations', International Conference on Document Analysis and Recognition, Bangalore, pp.169-172, 1999.

[4] Sural S., Das P.K., `A Two Step Algorithm and its Parallelisation for the Generation of Minimum Containing Rectangles for Document Image Segmentation', ICDAR, Bangalore, pp.173-176, 1999.

[5]Wang D., Srihari S. N.,`Classi_cation of Newspaper Image Blocks Using Texture Analysis',CVGIP, Vol. 47, pp. 327-352,1989

[6] Vishwanathan M., Nagy G., 'Characteristics of Digitized Images of Technical Articles', SPIE, Vol. 1, 661, pp. 6-17, 1992

[7] K. Jung, K. Kim and A.K. Jain, “Text Information Extraction in Images and Video: A Survey”, Pattern Recognition, Vol. 37, No. 5, pp. 977-997, May 2004.

[8] Song Mao, Azriel Rosenfeld, Tapas Kanungo Document structure analysis algorithms: a literature survey, Document Recognition and Retrieval X, Vol. 5010, No. 1., pp. 197-207, 2003.

[9] Nourbakhsh, F. Pati, P.B. Ramakrishnan, A.G. , Document Page Layout Analysis Using Harris Corner Points, Proceedings of ICISIP 2006.

[10] J.A. Montoya-Zegarra, N.J. Leite, R. Torres, "Rotation-Invariant and Scale-Invariant Steerable Pyramid Decomposition for Texture Image Retrieval", Brazilian Symposium on Computer Graphics and Image Processing,(1):121-128, Oct. 2007.

[11] Simoncelli, E.P., Freeman, W.T., The steerable pyramid: A flexible architecture for multi-scale derivative computation In: Proc. IEEE Second Internat. Conf. on Image Process. Washington, DC, pp. 444–447, 1995.

[12] L. O’Gorman, “The document spectrum for page layout analysis,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 15, no. 11, pp. 1162–1173, Nov. 1993.

[13] G. Nagy, S. Seth, and M. Viswanathan, “A prototype document image analysis system for technical journals,” Computer, vol. 7, no. 25, pp. 10–22, 1992.

[14] Refering to MediaTeam Document Database Sauvola J. and Kauniskangas H. MediaTeam Document Database II, a CD-ROM collection of document images, University of Oulu, Finland,1999.

[15] A. Antonacopoulos and S.Pletschacher and D. Bridson and C. Papadopoulos, ICDAR 2009 Page Segmentation Competition, in: 10th International Conference on Document Analysis and Recognition (ICDAR’09), Barcelona, Spain, July 2009.

[16] Otsu, N., "A Threshold Selection Method from Gray-Level Histograms," IEEE Transactions on Systems, Man, and Cybernetics, Vol. 9, No. 1, pp. 62-66,1979.