Marginal Noise Reduction in Historical Handwritten Documents - A Survey

Arpita Chakraborty
School of Information and Communication Technology
Griffith University
Gold Coast Campus, Australia
Email: arpita.chakraborty@griffithuni.edu.au

Michael Blumenstein
School of Software
University of Technology Sydney
Sydney, Australia
Email: Michael.Blumenstein@uts.edu.au

***Abstract*—This paper presents a survey on different approaches for removing the marginal noise from document images, and anlaysing the research challenges of those methods relating to handwritten historical datasets. In this survey, historical documents collected from Australian Archives and Libraries are introduced and the associated layout complexities of those document images are also described. Benchmarking other historical databases related to this work is also discussed. This survey discusses the difficulties and suitability of the state-of-the-art methods to remove marginal noise as well as preserving the text content from handwritten historical documents. This survey helps researchers to identify appropriate methods according to the associated marginal noise and also illustrates their drawbacks in order to make suggestions for developing approaches, which are more general and robust for any datasets.**

***Keywords*-Survey; Historical handwritten documents; Marginal noise removal**

I. Introduction

To ensure space-free preservation and open access of historical information, digitization (by scanning) of historical documents is the most conventional way used by libraries and archives throughout the world. Most of those historical scanned documents have lost the readability due to degradation. There are two types of degradation in document images: 1) physical degradation of the hardcopy documents during creation and/or storage and 2) degradation introduced by digitization [6]. Historical documents possess both types of degradation; either of them can reduce the performance of a document analysis system significantly. Processing those documents for transcription with the help of Optical Character Recognition (OCR) or similar applications is more challenging compared to a normal document. Several types of pre-processing techniques are required before targeting the goal of text recognition. A number of methods have been developed to reduce noise for non-historical printed or structured binary document images. This paper is an initial survey to introduce a review of the available methods in the literature and discuss the feasibility of applying those methods to remove the critical noise from historical handwritten documents. Apart from that, the additional challenges to remove marginal noise from historical handwritten document images are studied for each category of noise.

This paper is organized as follows: in section II, the characteristic of the marginal noise from various historical handwritten datasets (but non-exhaustive list of) are described. Section III describes state-of-the-art methods. Section IV presents conclusions.

II. Representation of Marginal Noise in Historical Datasets

A. Historical Datasets - some examples

The scanned historical document images from the Queensland State Archives, (*QSA*)[1], Australia and the State Library of Queensland, (*SLQ*)[2], Australia have lost their readability due to the degradation, old writing style, ink variation, etc. QSA holds records of many State departments, offices and corporations in the period of 1824−1908. The *QSA* dataset is comprised of multi-writer and multi-sized old manuscripts; the data contains tabular sheets, index, graphics (portraits and maps) with text in colour and binary formats. The *SLQ* dataset contains a significant portion of Queensland's documentary heritage, major reference and research collections.

The *Prosecution Project*[3], Griffith University is investigating the history of the criminal trial in Australia. There are records from various sources such as NSW State records, Queensland State Archive, State Records Office of Western Australia etc. Variation in the dataset is significant in terms of format, degradation, style, etc.

The *Parzival* database [2] is a 13th century multi-writer historical handwritten manuscript in German. The *Saint Gall* database [1] is a 9th century single writer historical handwritten manuscript in the Latin language. The *George Washington* dataset is written in English with ink on paper. These datasets are used in many recent research works.

University of Washington dataset (*UW-III*)[4] is the non-historical English/technical document image database produced by the Intelligent Systems Laboratory, at the University of Washington, Seattle, Washington, USA.

[1] http://www.archives.qld.gov.au/researchers/Pages/Default.aspx

[2] http://www.slq.qld.gov.au/resources/qld-history

[3] https://prosecutionproject.griffith.edu.au/

[4] http://isis-data.science.uva.nl/events/dlia//datasets/uwash3.html

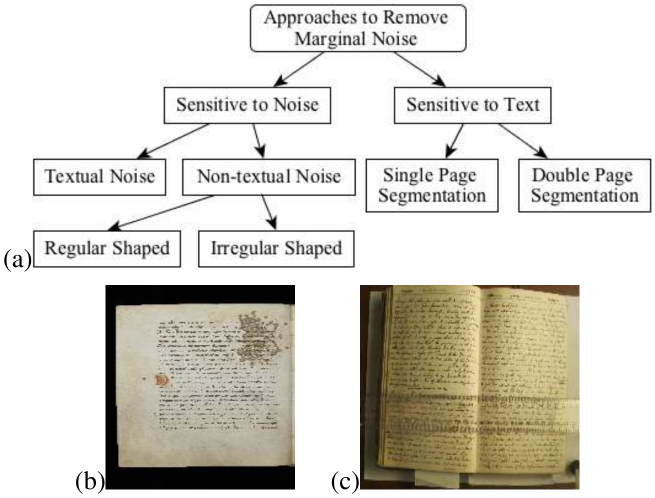

Figure 1. (a) Overview on approach and types of marginal noise. Examples of various types of noise in document images: (b) Regular shaped non-textual noise (*Saint Gall* dataset). (c) Textual and irregular shaped non-textual noise (*Prosecution Projectl* dataset).

B. Effect of Scanning

Marginal noise appears vertically and horizontally in a document image [7] which usually results from the scanning of thick or skew documents. More specifically, horizontal marginal noise is generated due to skew scanning and is located at the top or bottom border of a document image; while vertical marginal noise is caused by scanning thick and bounded documents and is located at the left or right border of a document image [9]. While scanning a thick document such as a book, the surface of the book is curved, when the scanner surface is flat. Hence such documents are scanned with non-uniform illumination. As an example, the gutter of a thick book can not touch the scanner and the scanning process is performed in changed illumination [9]. This process results in heavy darkness inside the margin. Usually, heavy darkness and shadow regions emerge in scanned documents due to this changing illumination. Thus scanning a book provides single page document with textual noise or double page documents.

C. Marginal Noise in Historical Documents

Marginal noise can be textual (text parts from a neighbouring page) or non-textual (black bars, speckles, etc.) with regular or irregular shapes and sizes [8] as shown in Fig. 1. This textual (Fig. 1(c)) and non-textual noise with either regular (Fig. 1(b)) or irregular shapes (Fig. 1(c)) are also observed in historical documents with further complexity. Marginal noise differs from page to page of the historical documents in terms of thickness (wide or thin), sharpness (faint or dark), shapes, length (continuous or broken), skew (slanted or straight), etc. There are additional spots or marks near the border such as punch-hole marks, torn pages, spots of water or ink, etc. within the multi-dimensional layout of historical document images. Variation in alignments of text lines and location as well as orientation of text contents are the critical and challenging issues for historical handwritten documents. Additionally, the handwritten comments/ notes or signs in different locations and orientations of the document images make the task more difficult for algorithms to remove marginal noise and preserve them along with the main text.

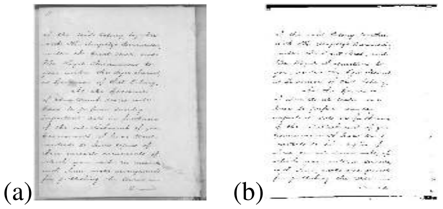

Figure 2. Effect of Binarization. (a) Marginal noise are observed at each margin of gray scale image from QSA dataset. (b) Left marginal noise is disappeared and only few dots are observed in the right margin of binary image of (a).

D. Effect of Format

Noise removal techniques for document images have mostly been developed for dealing with a binary format. In degraded historical documents, the gray-values of the pixels in the shadow region of marginal noise vary widely and are also much smaller than those of the non-shadow region. If we perform binarization by global thresholding, we lose the information as shown in Fig. 2(b). Hence conversion of degraded, age-affected manuscripts to a binary format has a high impact in information loss and thus it directly affects the accuracy of word/character segmentation and recognition. If significant amounts of marginal noise disappear after binarization, then the algorithm will fail to observe many critical situations and its performance cannot reach its highest performance.

In the literature, scanned gray-scale documents are binarized using various methods for minimizing information loss. These include adaptive binarization [3], in which the method is proposed to binarize historical documents used in experiments presented in [12], [17]. A local thresholding method is applied for the same purpose in [9].

E. Printed vs. Handwritten Documents

To implement and analyze a method, it is important to understand the motive and assumptions of that particular method. Sometimes the assumptions are developed based on the characteristics of the datasets. According to [6], the segmentation and recognition techniques applied for machine printed and handwritten text are significantly different. The difference between printed and handwritten documents is depicted in Table II-E.

Upon reviewing the literature, we found that the methods in [12], [15] perform based on the assumption that there is

Table I
PRINTED VS. HANDWRITTEN DOCUMENTS.

Characteristics	Printed documents	Handwritten documents
Page layout	Regular	Irregular [5]
Text line	Straight [5]	Curvilinear [5]
Gap between text and margin		
Presence	Always	Inconsistent
Size	Consistent	Inconsistent
Width	Consistent	Inconsistent
Text body alignment	Straight	Curvilinear
Character size	Single size	Multiple size

a consistent and minimum gap between the margin and text area in printed documents. This argument becomes invalid for the handwritten documents. For such documents, the text lines are curvilinear and their alignment is not straight. The location of the text area in handwritten documents also differs from page to page in multi-writter datasets. In Section III, the limitations of various methods are discussed based on the characteristics of document types as given in Table II-E.

F. Performance Evaluation

The performance evaluation of pre-processing algorithms is not very common in practice. One reason could be the cumbersome manual process to create the ground truth set of the original images. In most cases, the results are evaluated by visual inspection. In [17], a pixel based approach is applied to evaluate the performance of the method for marginal noise removal. This approach counts the pixels inside the frame of a manually created ground truth image (P_g) and the resulting image (P_R). The following equations are used to calculate the precision and recall as described in [17]:

$$\text{Precision} = \frac{T(P_g \cap P_R)}{T(P_R)} \quad \text{and} \quad \text{Recall} = \frac{T(P_g \cap P_R)}{T(P_g)}, \tag{1}$$

where $T(p)$ a function that counts the elements of set $T(p)$. Noise ratio and Page content removal techniques are also applied to measure the performance [8], [15]. The noise ratio of the document image is calculated to quantify the amount of border noise that remains in a document image. The purpose of measuring the percentage of ground-truth pixels removed from the image is to find the damage done to the actual page content area by the noise removal algorithm. These measures are given below:

$$\text{Noise ratio} = \frac{n_{pb}}{n_p} \quad \text{and} \quad \text{GT Removal} = \frac{n_p - n_c}{n_p}, \tag{2}$$

where n_{pb} is the number of foreground pixels outside the ground-truth page frame, n_p is the total number of foreground pixels in the actual page content area of a document image and n_c is the total number of foreground pixels in the cleaned image that matches pixels in the ground-truth image.

III. MARGINAL NOISE REMOVAL TECHNIQUES

Marginal noise removal techniques aim to preserve the text content while removing maximum amount of marginal noise of the document images. In the literature, few methods are implemented to remove such noise from historical handwritten documents; however there are many for printed and structured non-historical document images.

A review of the literature found that algorithms are categorized into two approaches: noise sensitive components and text sensitive components. The former one detects and deletes noisy components while the later one identifies the actual content area of the document. In this study, we elaborate the discussion based on these two approaches.

The Table II gives an overview of the various methods and their performance, data size and noise type. The performance of several methods presented in the 6th - 8th columns of Table II, are taken from the original paper cited and from [8]. The performance shows the efficiency to handle different types of marginal noise from various datasets. In the paper [8], the XY-cut algorithm along with six algorithms from both approaches are evaluated using 978 English printed documents from the University of Washington dataset (UW-III) and is shown in the 7th and 8th columns of Table II.

A. Noise Sensitive Components

The methods in this approach are developed based on the nature (textual and non-textual noise) of noise. It is observed in Table II that *Resolution reduction* [9] method performs well for both regular and irregular shaped non-textual noise whereas *Invading and non-invading algorithms* [10] work for irregular shaped non-textual noise. *Projection with Smearing* [12] and *Edge Density* [13] methods are applied to remove textual noise.

1. Non-Textual Noise

The *Resolution reduction* [9] method works in two steps: (i) Marginal noise blocks are detected by removing non-marginal noise blocks from the image using a reduction rate; this rate is equal to the average size of the characters in the image. Then the connected blocks are spilt horizontally and vertically by computing their run-lengths in the reduced image. The segmented blocks are identified as border noise components or non-border noise components based on their size, position, and neighbourhood. (ii) To delete noise regions, a polygonal boundary of each noise block is established and all the foreground pixels that lie within this boundary are removed from the original image. The block diagram of this method is shown in Fig. 3. This method shows better performance for preserving text content as well as removing noise (0.17% text content is removed and 71% marginal noise is removed, as shown in Table II). For gray-scale images, the marginal blocks are converted to a binary format by a local thresholding method.

Table II
OVERVIEW OF HANDLING MARGINAL NOISE BY VARIOUS ALGORITHMS.

Approach	Source	Methods	Suitable for			Datasize	Accuracy	Data set: UW-III [8]	
			Noise Type	Format	Data Type	# Images	(%)	Noise Ratio (%)	Page Contents Removal (%)
Noise Sensitive	[9]	Resolution Reduction	Regular & Irregular non-texual	Binary, Gray-scale	Printed	–	–	29.38	0.17
	[10]	Invading and non-invading	Irregular non-textual	Binary	Printed, Handwritten	20,000	95	–	–
	[12]	Projection with Smearing	Textual & Regular non-texual	Binary	Printed	1,705	78.82	8.38	6.96
	[13]	Edge Density	Textual & Regular non-texual	Binary	Printed	20	–	14.48	9.59
Text Sensitive	[14]	Page frame Detection	Textual & Non-textual	Binary	Printed	1,600	–	18.14	4.66
	[15]	Projection based cleanup	Textual & Regular non-texual	Binary	Printed	1,600	$70 - 20$	32.59	0.67
	[17]	Page frame Detection	Textual & Non-textual	Binary	Handwritten	458	99	–	–
	[18]	Page frame Detection	Regular Non-textual	Color	Handwritten	127	90	–	–

For handwritten documents, the character size varies a lot even in a single page. Estimating the average size of the characters in handwritten documents will be a challenge for this method.

The *Invading and non-invading algorithms* in [10] are able to remove irregular shaped non-textual marginal noise from binary document images for (i) noisy border merges with a document, (ii) noisy border towards the document, (iii) narrow irregular vertical lines and (iv) islands with black pixels. This method shows better performance compared with the available commercial tools. The method in [11] includes a pre-processing step on the improved [10] algorithm and gains better performance with speeding up the flood-fill process.

This flood-fill method [10] detects the threshold moving from each pixel from the list of black pixels to the left to right until it reaches a white pixel in binary images. This technique to choose a threshold point is hard to apply for a gray-scale image. This may cause either a loss of text information or a big amount of noise will remain in the gray-scale document image.

2. Textual Noise

The *Projection with Smearing* [12] method works in separate steps to remove textual and non-textual noise. First, this method uses the run-length smearing algorithm to smooth binary images. Connected component labelling is then performed. The limits of text regions are computed horizontally and vertically using horizontal and vertical projection profile. In the cleanup stage, all the black pixels that belong to a connected component with at least one pixel lying outside the detected page content limit are transformed to white. Similarly, textual noise is detected and removed. After visual checking, noise is correctly removed for 1344 images (78,82% of testing set).

The method is proposed under the assumption that the marginal noise is not too close to the actual text content. If so, there is a scope to loose a significant amount of the text region. For handwritten documents, it is a very common to get the text content very close to any border.

The key idea of the *Edge Density* method [13] is that text areas have a low density of edges while border noise areas have a high edge density. The algorithm works in three steps: (i) edge detection using the sobel operator, (ii) marginal noise detection from the projection profile using critical density and (iii) marginal noise deletion by a coarse-to-fine method. From the statistics shown in Table II, we can conclude that a large amount of page content is removed as noise, whilst the noise ratio is still high.

This method searches the single sharp peak near the margin. To apply this method for historical handwritten documents, the following issues could arise: (i) The noise could be as wide as the document image (Fig. 1(c)) and (ii) there could be more than one peak for the degraded document image. The ruled lines throughout the document image can also be detected as sharp peaks in the projection profiles.

B. Text Sensitive Components

In this approach, the main focus is to identify the page frame of the document images using various properties of the text content. The page frame is a small rectangle that encloses all the foreground elements of the document image. According to [7], the performance of the algorithms from

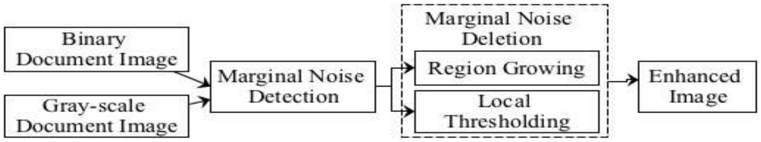

Figure 3. The block diagram for removing non-textual noise. Source: [9].

this group is better than the former one because searching for text patterns is much easier than searching for the features of noise in any document. Most of the methods [14], [15], [18] are used to segment the text content for single page images. Page frame detection in [17] is developed for double page segmentation.

1. Single Page Segmentation

In [14], the proposed method works in two steps: (i) A geometric model is built for the page frame of a scanned document. (ii) A geometric matching method is used to find the globally optimal page frame with respect to a defined quality function. This method can also detect the page frame when the noise overlaps some regions of the page content area. The performance is not affected even if there is no whitespace between the marginal noise and the page frame. Several error measures are performed based on area overlap, connected component classification, and ground-truth zone detection accuracy for determining the accuracy of the algorithm. The major source of errors is missing isolated page numbers. From Table II, this method is not able to keep most of the page content of the structured document when the noise ratio is high.

This method requires prior extraction of text lines and zones from the document images and this process makes the process makes it slow and hard to implement. In historical handwritten document images, the orientation and location of the side notes are distributed in an unstructured way and this will be a challenge for this method.

The *Projection based cleanup* [15] method works in three steps: (i) A black filter is used to select the large black regions at the margins that are bigger than a pre-defined threshold area; (ii) All connected components close to the border are detected as noise and so removed from the image. An appropriate value for the threshold is selected which depends on prior knowledge. (iii) A white filter is applied which extracts features similar to the black filter and then removes everything up to the border if it finds a large white block. According to Table II, this method is able to keep most of the page content while the noise ratio is high.

This method works on the assumption that there are always white spaces in between the border and actual page contents in a document. Such assumption for handwritten documents is void, as the location of the text content as well as the alignments and space with the margin, differ in significant manner.

The *Page segmentation* method in [18] works in three steps to remove regular-shaped non-textual noise: (i) The feature vector is extracted concatenating features from color (Variance, Smoothness and Laplacian), coordinates and texture (Local Binary Pattern and Gabor Dominant Orientation Histogram) of the document image; (ii) Optimal feature subset is then selected using the Fast Correlation-Based Filter to reduce the dimensionality of the feature space and (iii) Support Vector Machines are applied for classification. As a post-processing step, pixels are classified into four classes: periphery, background, text and decoration. This method is evaluated on three historical handwritten documents and achieves approximately 91% to 98% accuracy on *Parzival*, *Saint Gall* and *George Washington* datasets.

2. Double Page Segmentation

In [17], the proposed method detects the optimal page frames of double-page document images based on the white run projections. For this experiment, a double-page document image is identified if the width length is higher than the height of the page document. After preprocessing, all noisy small components having a height and width less than ten times the average character height are removed from the image. The vertical and horizontal zones are then detected analyzing white run projections. Three cases are considered for vertical zone detection from document images: two vertical zones, one vertical zone with an empty page or a little text and more than two vertical zones. The block diagram depicts the steps of segmentation shown in the Fig. 4.

This method can remove textual and non-textual noise only for structured double-page documents such as journals or newspaper articles. The method is approximately 99% accurate in removing borders without cropping page content from historical document images. The performance drops for multiple columns with a complex layout. It is also observed in the resulting figures that the method fails to preserve the page content for closed zones or those where there are no gaps between two zones of the document images. In this method, there are 7 parameters to fix manually.

IV. Conclusion

The literature on page segmentation for historical handwritten documents is limited, although there are many for non-historical documents. This paper describes various methods for marginal noise reduction, and explains those

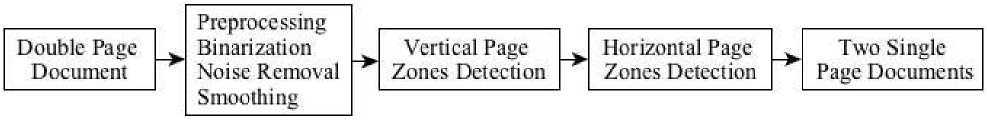
Double Page Document
Preprocessing
Binarization
Noise Removal
Smoothing
Vertical Page Zones Detection
Horizontal Page Zones Detection
Two Single Page Documents

Figure 4. The block diagram for double page segmentation. Source: [17].

as applied to handwritten document images. This survey summarizes the state-of-the-art and identifies the gaps and difficulties for implementing solutions for historical handwritten document images. A comparison of methods on the basis of experimental results on historical handwritten document images could be considered as a scope for future work.

References

[1] A. Fischer, V. Frinken, A. Forns, and H. Bunke, *Transcription Alignment of Latin Manuscripts using Hidden Markov Models*, in Proc. 1st Int. Workshop on Historical Document Imaging and Processing, 29-36, 2011.

[2] A. Fischer, A. Keller, V. Frinken, and H. Bunke, *Lexicon-Free Handwritten Word Spotting Using Character HMMs*, in Pattern Recognition Letters, 33(7): 934-942, 2012.

[3] B. Gatos, I. Pratikakis and S. J. Perantonis, *Adaptive Degraded Document Image Binarization*. Pattern Recognition, 39, 317-327, 2006

[4] Z. Hadjadj, A. Meziane, M. Cheriet and Y. Cherfa, *An Active Contour Based Method for Image Binarization: Application to degraded historical document images.* 14th ICFHR. 2014.

[5] L. Yi, Y. Zheng and D. Doermann and S. Jaeger, *Script-Independent Text Line Segmentation in Freestyle Handwritten Documents*, IEEE Trans Pattern Analysis and Machine Intelligence. 30(8): 1313 - 1329, 2008.

[6] Y. Zheng, H. Li and D. Doermann, *Machine Printed Text and Handwriting Identification in Noisy Document Images*, IEEE Trans Pattern Analysis and Machine Intelligence. 26(3): 337 - 353, 2003.

[7] A. Farahmand, A. Sarrafzadeh and J. Shanbehzadeh, *Document Image Noises and Removal Methods*, International Multi Conference of Engineers & Computer Scientists, 1(1), 2013.

[8] F. Shafait and T.M. Breuel, *The effect of border noise on the performance of projection-based page segmentation methods.* IEEE Trans Pattern Analysis and Machine Intelligence. 33(4): 846-51, 2011.

[9] K.-C. Fan, Y.-K. Wang, and T.-R. Lay, *Marginal noise removal of document images*, Pattern Recognition Society, Elsevier Science Ltd., 2593-2611, 2002.

[10] B. T. Avila and R. D. Lins, R.D, *A New Algorithm for Removing Noisy Borders from Monochromatic Documents*, Proc. of ACM-SAC 2004, Cyprus, ACM Press,1219-1225, 2004.

[11] A. A. Formiga and R. D. Lins, *Efficient Removal of Noisy Borders of Monochromatic Documents.*, Lecture Notes in Computer Science. ed.Berlin: Springer Berlin Heidelberg, 5627: 158-167, 2009. .

[12] N. Stamatopoulos, B. Gatos, and A. Kesidis, *Automatic borders detection of camera document images*, in 2nd Int. Workshop on Camera-Based Document Analysis and Recognition, Curitiba, Brazil, Sep. 2007, pp. 71- 78.

[13] W. Peerawit and A. Kawtrakul, *Marginal Noise Removal from Document Images Using Edge Density*, Proceedings of Fourth Information and Computer Eng. Postgraduate Workshop, January 2004.

[14] F. Shafait, J. van Beusekom, D. Keysers, and T.M. Breuel, *Document cleanup using page frame detection*, Int. Jour. on Document Analysis and Recognition, 11(2):81- 96, 2008.

[15] F. Shafait and T. M. Breuel, *A simple and effective approach for border noise removal from document images*, in 13th IEEE Int. Multi-topic Conference, Islamabad, Pakistan, Dec 2009.

[16] L. Gorman, *The Document Spectrum for Page Layout Analysis*, IEEE Trans. Pattern Analysis and Machine Intelligence, 15(11): 1162-1173, 1993.

[17] N. Stamatopoulos, B. Gatos and T. Georgiou, *Page Frame Detection for Double Page Document Images*, in Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, Boston, MA, USA, 2010, pp. 401 - 408.

[18] K. Chen, H. Wei, J. Hennebert, R. Ingold and M. Liwicky, *Page Segmentation for Historical Handwritten Document Images using color and texture fatures*, 14th International Conference on Frontiers in Handwriting Recognition, 2014.

[19] F. Shafait, D. Keysers and T. M. Breuel, *Performance Evaluation and Benchmarking of Six-Page Segmentation Algorithms*, IEEE Trans Pattern Analysis and Machine Intelligence. 30(6): 941 - 954, 2008.