A Comprehensive Evaluation Methodology for Noisy Historical Document Recognition Techniques

Nikolaos Stamatopoulos[1] Georgios Louloudis[1,2] Basilis Gatos[1]

[1]Computational Intelligence Laboratory, Institute of Informatics and Telecommunications,
NCSR "Demokritos", GR-15310
Agia Paraskevi, Athens, Greece
{nstam,bgat}@iit.demokritos.gr

[2]Department of Informatics and Telecommunications,
University of Athens, Greece
louloud@mm.di.uoa.gr

ABSTRACT
In this paper, we propose a new comprehensive methodology in order to evaluate the performance of noisy historical document recognition techniques. We aim to evaluate not only the final noisy recognition result but also the main intermediate stages of text line, word and character segmentation. For this purpose, we efficiently create the text line, word and character segmentation ground truth guided by the transcription of the historical documents. The proposed methodology consists of (i) a semi-automatic procedure in order to detect the text line, word and character segmentation ground truth regions making use of the correct document transcription, (ii) calculation of proper evaluation metrics in order to measure the performance of the final OCR result as well as of the intermediate segmentation stages. The semi-automatic procedure for detecting the ground truth regions has been evaluated and proved efficient and time saving. Experimental results prove that using the proposed technique, the percentage of time saved for the text line, word and character segmentation ground truth creation is more than 90%. An analytic experiment using a commercial OCR engine applied to a historical book is also presented.

Categories and Subject Descriptors
I.4.6 [Image Processing]: Segmentation; I.4.9 [Image Processing]: Applications; I.5.4 [Pattern Recognition]: Applications---Text Processing; I.7.5 [Document and Text Processing]: Document Capture---Document Analysis.

General Terms
Evaluation, Performance, Experimentation

Keywords
Document Image Processing, Historical Document Processing and Recognition, Evaluation, OCR, Transcript Mapping, Segmentation, Text Line Segmentation, Word Segmentation.

1. INTRODUCTION
Recognition of historical documents is essential for quick and efficient content exploitation of the valuable historical collections that are part of our cultural heritage. Several factors such as low paper quality, dense and arbitrary layout, low print contrast, typesetting imperfections, lack of standard alphabets and fonts do not permit the application of conversional recognition techniques to historical documents. Due to these reasons, recognition of historical documents is one of the most challenging tasks in document image processing. To this end, several techniques that focus on historical document processing and recognition have been proposed recently in the literature. Most of these works focus on the unique characteristics of the corresponding historical document such as content and writing style. In [7, 11] OCR systems were developed respectively for the recognition of characters used in the Christian Orthodox Church Music notation. In [15], an approach for the recognition of Early Christian Greek manuscripts based on the detection of open and closed cavities of the skeletonised characters is presented. The basic characteristic of these documents is that there is no space between two consecutive words. In [4] an open-source programming framework is introduced for building systems that extract information from digitized historical documents empowering the document experts themselves to develop systems with reduced effort. In [19], a complete OCR methodology for recognizing historical documents, either printed or handwritten without any knowledge of the font, is presented. It consists of a pre-processing step, a top-down segmentation step as well as a clustering scheme in order to group characters of similar shape. A segmentation-free approach is followed in [17, 12, 2] where line and word segmentation is used for creating an index based on word matching. In [9], a word spotting technique based on combining synthetic data and user feed-back for keyword searching in historical printed documents is described.

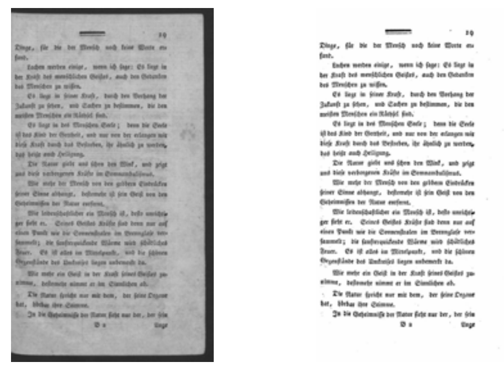
Text resulting from historical document recognition is noisy since it corresponds to low recognition rates. In most cases, historical document recognition systems produce a recognition result that is evaluated in terms of character accuracy at the levels of 90% - 95% [15, 19]. One of the reasons for this is the fact that several errors are introduced during the segmentation phase of historical documents. Several problems such as low quality, complex, dense and irregular layouts, noise between characters, ink diffusion and local skew seriously affect the segmentation and, consequently, the recognition accuracy. To this end, it is imperative to have an objective evaluation which will account for the performance not only of the final recognition result but also of the involved document image segmentation stages. This will help to detect the processing stages that introduce the majority of errors and consequently guide future efforts and research for the development of an improved recognition module.

Motivated by the aforementioned goal, in this paper we propose a new comprehensive methodology in order to evaluate not only the final noisy recognition result but also the main intermediate stages of segmentation of historical documents. The main novelties of the proposed methodology comprise (i) an efficient and time-saving semi-automatic methodology for the creation of the text line, word and character segmentation ground truth taking into consideration the transcription for each segmentation level; (ii) a comprehensive evaluation methodology in order to evaluate not only the final noisy recognition result but also the main intermediate stages of text line, word and character segmentation of historical documents. Guided by the transcription of the historical documents we efficiently create the text line, word and character segmentation ground truth which is used in order to produce a comprehensive evaluation report. The paper is organized as follows: In Section 2, we describe the proposed methodology for the evaluation of historical document recognition techniques focusing on the distinct steps we follow. In Section 3, the semi-automatic procedure for detecting the ground truth regions is evaluated while an analytic experiment using a commercial OCR engine applied to a historical book is presented. Finally, in Section 4 conclusions are drawn.

2. PROPOSED EVALUATION METHODOLOGY

According to the proposed methodology, in order to evaluate a historical document recognition technique we need (i) the document's transcription which is the correct ASCII text of the document and (ii) the recognition result as well as the text line, word and character segmentation result of the OCR engine under evaluation. First, we process the initial historical document image in order to obtain a binary image output I of accepted quality since our evaluation methodology mainly focuses on recognition and segmentation issues. We use the binarization methodology of [6] as well as the border removal algorithm of [18] with suitable parameter tuning and with the help of visual inspection in order to achieve a binary image result I of accepted quality that efficiently preserves text areas (see Figure 1).

We assume that the transcription includes the correct text line break information. For each page, the transcription is first processed by a simple text parsing module in order to detect the number of text lines as well as the number of words and characters for every text line. This information is used in a transcript mapping module in order to efficiently create the text line, word and character segmentation ground truth. Transcript mapping (or text alignment) techniques are used in order to map the correct text information to a segmentation result produced automatically. Usually, these techniques are very useful in order to automatically create benchmarking data sets. In [21], automatic segmentation of cursive handwritten text lines is achieved using the transcriptions of the text lines and a hidden Markov model (HMM) based recognition system. In [10], an algorithm based on dynamic time warping (DTW) is proposed for a word by word alignment of handwritten documents with their (ASCII) transcripts. In our approach, in order to create the text line segmentation ground truth we first use a Hough transform based methodology guided by the number of the text lines indicated at the previous text parsing stage. Concerning the word and character segmentation ground truth, we first use a gap classification technique constrained by the number of the words and characters for every text line indicated at the text parsing stage. For the creation of the final text line, word and character segmentation ground truth we also involve a user guided correction module. As it will be demonstrated in Section 3, only a small number of segmentation results needs correction since the proposed automatic transcript mapping technique has been proved efficient and time saving.



(a) (b)

Figure 1. (a) Initial historical document image; (b) Binary image output after binarization and border removal.

The evaluation of the recognition engine is based on comparing the text recognition result with the transcription as well as on comparing the text line, word and character segmentation result with the corresponding segmentation ground truth. Text result evaluation is based on the edit distance calculation [8] while segmentation evaluation is based on pixel based area recall and precision calculation [5, 16]. The flowchart of the proposed evaluation methodology is demonstrated in Figure 2 while all involved steps are detailed in this section.

2.1 Transcription Text Parsing

Transcription contains useful information that can be used in order to correctly segment a document image into text lines, words and characters. We assume that we have a transcription file per document image as well as that the transcription includes line break information. By using a simple text parser, we calculate the number of text lines NL appearing in the document image as well as the number of words NW_i and characters NC_i of every text line i.

2.2 Transcript Mapping

Guided by the transcription information extracted in the previous section we efficiently create the text line, word and character segmentation result in order to facilitate the segmentation ground truth construction.

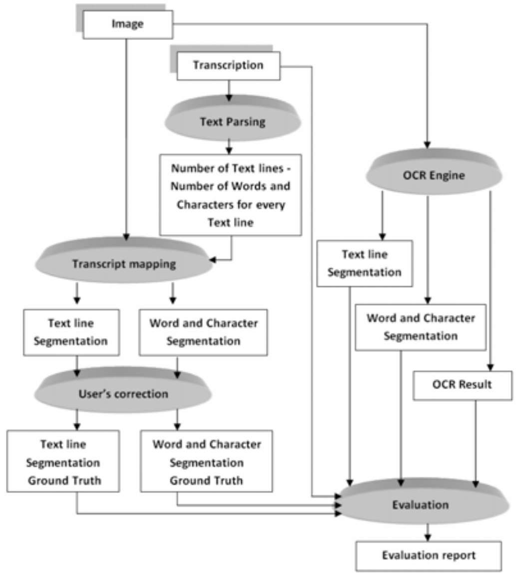


Figure 2. The flowchart of the proposed evaluation methodology for historical document recognition techniques.

2.2.1 *Text line segmentation*

The methodology for the segmentation of a document image into text lines is a modification of the methodology described in [13] which takes into consideration the number of lines *NL*. It includes three stages: (i) pre-processing, (ii) Hough transform mapping and (iii) post-processing.

2.2.1.1 *Pre-processing*

At this stage we apply a connected component rule based algorithm in order to exclude several non-text elements. The pre-processing stage consists of the following steps. In the first step, the connected components [3] (CCs) of the binary image *I* are extracted. Then, the average character height *AH* of the whole document image is calculated based on the average height of all CCs. The next step concerns the application of a horizontal Run Length Smoothing Algorithm (RLSA) [20] with a threshold value of *0.4*AH*. Then, the connected components on the smoothed image are calculated and areas that satisfy the following condition on the smoothed image are removed from image *I*:

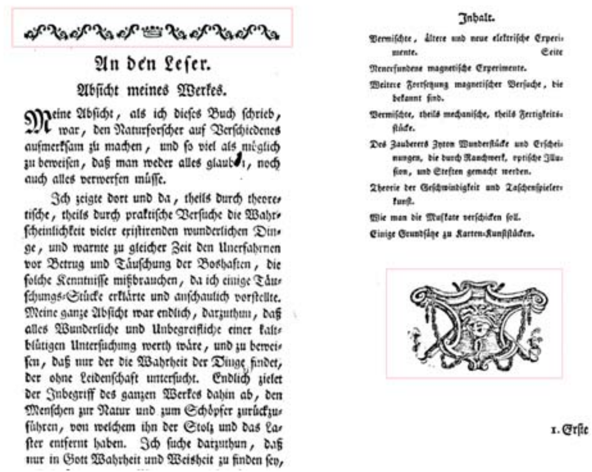
$$(CS_w > 0.6 I_w) \text{ OR } (CS_h > 5AH) \tag{1}$$

where I_w corresponds to the width of the document image *I* and CS_w, CS_h correspond to the width and height of a connected component on the smoothed image, respectively. Figure 3 shows two examples with non-text areas that are removed based on the condition of eq. (1).

Furthermore, in the pre-processing stage we exclude horizontal graphical line elements. We remove all connected components of image *I* which satisfy the following condition:

$$\left(\frac{n_b}{CI_w\, CI_h} > 0.5\right) \text{ AND } \left(\frac{CI_w}{CI_h} > 3\right) \text{ AND } (CI_w > 3AH) \tag{2}$$

where n_b, CI_w and CI_h are the number of black pixels, the width and the height of the connected component respectively. With this condition we eliminate black graphical lines as the ones appearing in Figure 4.



An den Leser.

Absicht meines Werkes.

Meine Absicht, als ich dieses Buch schrieb, war, den Naturforscher auf Verschiedenes aufmerksam zu machen, und so viel als möglich zu beweisen, daß man weder alles glaub[illegible], noch auch alles verwerfen müsse.

Ich zeigte dort und da, theils durch theoretische, theils durch praktische Versuche die Wahrscheinlichkeit vieler existirenden wunderlichen Dinge, und warnte zu gleicher Zeit den Unerfahrnen vor Betrug und Täuschung der Boshaften, die solche Kenntnisse mißbrauchen, da ich einige Täuschungs-Stücke erklärte und anschaulich vorstellte. Meine ganze Absicht war endlich, darzuthun, daß alles Wunderliche und Unbegreifliche einer kaltblütigen Untersuchung werth wäre, und zu beweisen, daß nur der die Wahrheit der Dinge findet, der ohne Leidenschaft untersucht. Endlich zielet der Inbegriff des ganzen Werkes dahin ab, den Menschen zur Natur und zum Schöpfer zurückzuführen, von welchem ihn der Stolz und das Laster entfernt haben. Ich suche darzuthun, daß nur in Gott Wahrheit und Weisheit zu finden sey,

Inhalt.

Vermischte, ältere und neue elektrische Experimente. Seite

Neuerfundene magnetische Experimente.

Weitere Fortsetzung magnetischer Versuche, die bekannt sind.

Vermischte, theils mechanische, theils Fertigkeitsstücke.

Des Zauberers Jyton Wunderstücke und Erscheinungen, die durch Rauchwerk, optische Illusion, und Stefsten gemacht werden.

Theorie der Geschwindigkeit und Taschenspielerkunst.

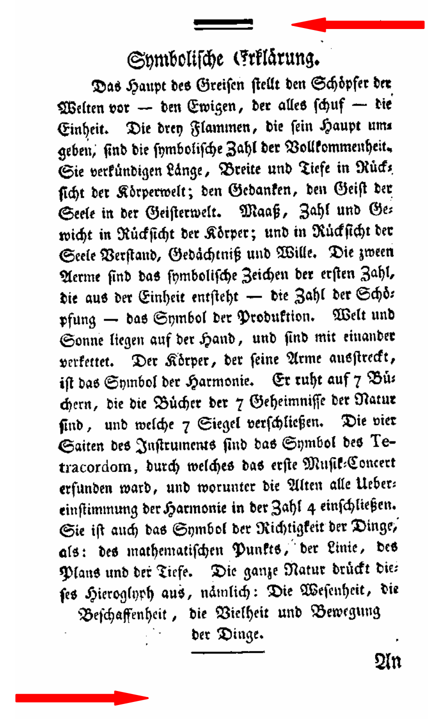
Wie man die Muskate verschicken soll.

Einige Grundsätze zu Karten-Kunststücken.

I. Erste

(a) (b)

Figure 3. Two examples with non-text areas (indicated by a rectangular box) that are removed during the pre-processing stage.



Symbolische Erklärung.

Das Haupt des Greisen stellt den Schöpfer der Welten vor — den Ewigen, der alles schuf — die Einheit. Die drey Flammen, die sein Haupt umgeben, sind die symbolische Zahl der Vollkommenheit. Sie verkündigen Länge, Breite und Tiefe in Rücksicht der Körperwelt; den Gedanken, den Geist der Seele in der Geisterwelt. Maaß, Zahl und Gewicht in Rücksicht der Körper; und in Rücksicht der Seele Verstand, Gedächtniß und Wille. Die zween Aerme sind das symbolische Zeichen der ersten Zahl, die aus der Einheit entsteht — die Zahl der Schöpfung — das Symbol der Produktion. Welt und Sonne liegen auf der Hand, und sind mit einander verkettet. Der Körper, der seine Arme ausstreckt, ist das Symbol der Harmonie. Er ruht auf 7 Büchern, die die Bücher der 7 Geheimnisse der Natur sind, und welche 7 Siegel verschließen. Die vier Saiten des Instruments sind das Symbol des Tetracordom, durch welches das erste Musik-Concert erfunden ward, und worunter die Alten alle Uebereinstimmung der Harmonie in der Zahl 4 einschließen. Sie ist auch das Symbol der Richtigkeit der Dinge, als: des mathematischen Punkts, der Linie, des Plans und der Tiefe. Die ganze Natur drückt dieses Hieroglyph aus, nämlich: Die Wesenheit, die Beschaffenheit, die Vielheit und Bewegung der Dinge.

An

Figure 4. An example of graphical lines (indicated by arrows) that are removed during the pre-processing stage.

2.2.1.2 Hough Transform Mapping

The main stage of the text line segmentation methodology is the application of the Hough transform on a set of points (see [13]).

The Hough transform is a line to point transformation from the Cartesian space to the Polar coordinate space. A line in the Cartesian coordinate space is described by the equation:

$$x\cos(\theta) + y\sin(\theta) = p \tag{3}$$

It is easily observed that the line in the Cartesian space is represented by a point in the Polar coordinate space whose coordinates are p and θ. Every point which votes, corresponds to a set of cells in the accumulator array of the (p,θ) domain. To construct the Hough domain the resolution along θ direction was set to 1 degree letting θ take values in the range 85 to 95 degrees and the resolution along p direction was set to $0.2*AH$.

After the computation of the accumulator array we proceed to the following procedure: We detect the cell (p_i, θ_i) having the maximum contribution and we assign to the text line (p_i, θ_i) all points that vote in the area $(p_i\text{-}5, \theta_i)..(p_i+5, \theta_i)$. The difference from the methodology presented in [13] is that instead of calculating the cell (p_i, θ_i) having the maximum contribution until a stopping criterion is met, we take into account that we expect NL maximum cells (p_i, θ_i).

2.2.1.3 Post-processing

The post-processing procedure consists of two steps. At the first step, a merging technique over the result of the Hough transform is applied to correct some false alarms. This stage may reduce the number of text lines detected by the Hough transform. In the second stage, connected components that were not clustered to any line are checked to see whether they create a new line that the Hough transform did not reveal. We force the algorithm to create as many text lines as are needed in order to reach the desired value NL. After the creation of the final set of lines, all unclassified components are grouped to the closest line. There are cases where although the algorithm is forced to produce NL number of text lines, the intermediate steps can produce a number of lines smaller than this value. The reason for this is that although the number of text lines is less than NL, there do not exist any unclassified components for the creation of new text lines.

A text line segmentation result of a document image portion is presented in Figure 5.

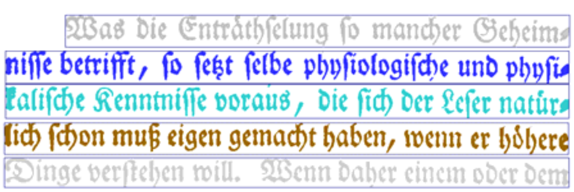


Figure 5. Text line segmentation result of a document image portion.

2.2.2 Word segmentation

The word segmentation procedure is divided into two steps. The first step deals with the computation of the distances of adjacent components in the text line image and the second step is concerned with the classification of the previously computed distances as either inter-word distances or inter-character distances. The methodology that was used for this task is a modification of the methodology presented in [14].

2.2.2.1 Distance Computation

In order to calculate the distance of adjacent components in the text line image, a pre-processing procedure is applied. The pre-processing procedure concerns the correction of the skew angle of the text line image. The computation of the gap metric is considered not on the connected components (CCs) but on the overlapped components (OCs), where an OC is defined as a set of CCs whose projection profiles overlap in the vertical direction.

The Euclidean distance is used as distance between two adjacent overlapped components (OCs). The Euclidean distance between two adjacent overlapped components is defined as the minimum Euclidean distance among the Euclidean distances of all pairs of points of the two adjacent overlapped components. For the calculation of the Euclidean distance we apply a fast scheme that takes into consideration only a subset of the pixels of the left and right OCs instead of the whole number of black pixels. In order to define the subset of pixels of the left OC, we include in this subset the rightmost black pixel of every scanline. The subset of pixels for the right OC is defined by including the leftmost black pixel of every scanline. Finally, the Euclidean distance of the two OCs is defined as the minimum of the Euclidean distances of all pairs of pixels.

2.2.2.2 Gap Classification

For the gap classification we use a local threshold for every text line of the image. All distances above this threshold are considered as inter-word gaps whereas all distances below this threshold are considered as intra-word gaps. In order to calculate this threshold on a text line i, we use the number of words NW_i of the particular text line which is calculated from the transcription (Section 2.1) as follows:

Let L be the number of the overlapped components of the text line image. The total number of distances computed is $L\text{-}1$. We define these distances as d_j, j=1..L-1. We sort the distances d_j in descending order. Defining the first distance as candidate to be the segmentation threshold, we make the segmentation and count the number of words that are produced. If the number or words produced is equal or larger than the value NW_i then this distance is the desired threshold for the particular text line. Otherwise, the next distance in the sorted list is considered as a threshold and the above described procedure is repeated until one distance meets the requirement.

A word segmentation result of a document image portion is presented in Figure 6. In Figure 7 we present an example of word segmentation results using different segmentation thresholds. Since the expected number of words NW_i is 4, we select the result of Figure 7(d).

Figure 6. Word segmentation result of a document image portion.

(a)

(b)

(c)

(d)

Figure 7. An example of word segmentation results using different segmentation thresholds.

2.2.3 Character segmentation

The character segmentation module uses a local threshold for every text line image in order to detect the character regions. The methodology for the calculation of the local threshold is similar to the one described in Section 2.2.2. The main differences are: (i) the segmentation procedure works on connected components instead of overlapped components and (ii) the number of regions that need to be calculated is NC_i instead of NW_i. A character segmentation result of a document image portion is presented in Figure 8.

Figure 8. Character segmentation result of a document image portion.

2.3 Correction of Segmentation Results

Once the text lines, the words and the characters have been detected, making use of the correct document transcription (Section 2.2), the user corrects possible segmentation errors in order to produce the final segmentation ground truth.

The user is provided with an appropriate tool to handle segmentation errors. Figure 9 is a screenshot of the developed tool. The tool enables the user to perform a few tasks to finalize the ground truth regions such as editing, inserting or deleting segmentation regions. As it is demonstrated in Section 3, only a small number of segmentation results needs correction since the proposed automatic transcript mapping technique has been proved efficient and time saving. Among segmentation results that need correction, there are cases of ornamental characters which get split in the segmentation procedure due to their large height (Fig. 10). The correction made by the user is the assignment of the ornamental character to a single text line and word. After this procedure, the text line, word and character segmentation ground truth regions are defined in order to be used for evaluation.

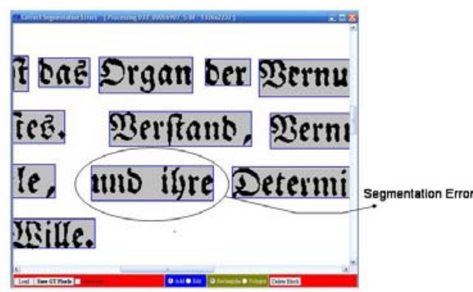
2.4 Evaluation Metrics

2.4.1 Segmentation Evaluation Metrics

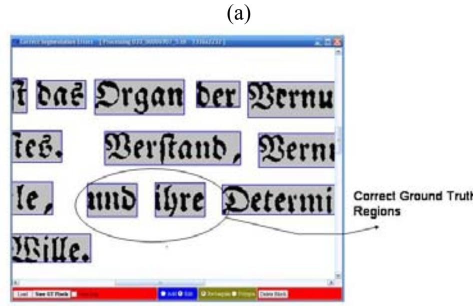
The performance evaluation of text line, word and character segmentation is based on counting the number of the matches between the entities detected by the segmentation algorithm and the ground truth [5, 16]. It is a pixel based method and takes into account only the black pixels of the entities ignoring the white pixels of background.

Let I be the set of all image black points, G_i the set of all black points inside the i ground truth region, R_j the set of all black points inside the j result region, $T(s)$ a function that counts the elements of set s. Table *MatchScore(i,j)* represents the matching results of the i ground truth region and the j result region as follows:

$$MatchScore(i,j) = \frac{T(G_i \cap R_i \cap I)}{T((G_i \cap R_i) \cup I)} \quad (4)$$



(a)



(b)

Figure 9. (a) Word segmentation result; (b) Ground truth regions after user's intervention.

The performance evaluator searches within the *MatchScore* Table for pairs of one-to-one matches. We call a pair one-to-one match (*o2o*) if the matching score for this pair is equal to or above the evaluator's acceptance threshold T_a of the total region. If *N* is the count of ground truth regions and M the count of result regions we calculate the detection rate (*DR*) and recognition accuracy (*RA*) as follows:

$$DR = \frac{\#o2o}{N} \tag{5}$$

$$RA = \frac{\#o2o}{M} \tag{6}$$

A performance metric FM can be extracted if we combine the values of detection rate and recognition accuracy:

$$FM = \frac{2 \cdot DR \cdot RA}{DR + RA} \tag{7}$$

2.4.2 OCR Evaluation Metrics

The optical character recognition system will be evaluated on character level using the Character Accuracy metric [8]. In order to define this metric we need to define when an error occurs. We count an error for each character insertion, deletion or substitution that is required to correct the text generated by the optical character recognition system. The number of errors is defined as the minimum number of edit operations (character insertions, deletions and substitutions) needed to fully correct the text. Thus, Character Accuracy is defined as the ratio of the number of correct characters (number of characters in the correct document transcription minus number of errors) to the total number of characters in the correct document transcription:

$$Character\ Accuracy = \frac{\#characters - \#errors}{\#characters} \tag{8}$$

3. EXPERIMENTAL RESULTS

The evaluation methodology presented in the previous section was tested on a part of a historical book from Eckartshausen which was published on 1788 and is owned by the Bavarian State Library[1] consisting of 94 document images that contained 2647 text lines, 18575 words and 116887 characters.

First, we evaluate the proposed technique in order to detect the ground truth regions before the user's intervention. For this reason, we manually marked the correct line and word segments in the set of 94 images as well as the correct character segments in 18 representative images. Table 1 shows the detailed evaluation results where the acceptance threshold is set to $T_a = 95\%$ for text line segmentation and $T_a = 90\%$ for word and character segmentation.

[1] Carl von Eckartshausen, "Aufschlüsse zur Magie aus geprüften Erfahrungen über verborgene philosophische Wissenschaften und verdeckte Geheimnisse der Natur", Bavarian State Library, 1778.

Table 1. Evaluation results for the proposed segmentation technique making use of the correct document transcription.

	N	M	o2o	DR	RA	FM
Text Lines	2647	2646	2618	98,9%	98,9%	**98,9%**
Words	18575	18585	18248	98,2%	98,1%	**98,1%**
Characters	14667	14755	13291	90,6%	90,0%	**90,2%**

It can be observed from the *o2o* matches that the proposed technique fails to correctly detect 29 text lines out of 2647, 327 words out of 18575 and 1376 characters out of 14667. All these segments should be manually corrected at the next step of the procedure (see Section 2.3).

We recorded that for the manual creation of the text line, word and character ground truth, for one document image, the user needs 420, 900 and 5400 seconds in average, respectively. Concerning the proposed methodology, the average time for visually checking the generated text line, word and character segmentation result is about 20, 40 and 80 seconds per image, respectively. The average time needed to correct a single segmentation error (either text line, word and character) using the tool described in Section 2.3 is about 5 seconds. A user needs 20+(5*29/94)≈22 seconds for the correction of the text line segmentation results per image, 40+(5*327/94)≈58 seconds for the correction of the word segmentation results per image and 80+(5*1376/18)≈463 seconds for the correction of the character segmentation results per image. Taking into account the abovementioned rationale, we can state that the percentage of time saved for the text line, word and character segmentation ground truth creation is about 92%.

Figures 10, 11 and 12 demonstrate representative problems that are encountered during the text line, word and character segmentation procedure. Ornamental characters and graphical illustration which couldn't be removed at the pre-processing step are the main causes of segmentation errors. Also, there are cases that punctuation marks are detected as single words and this causes erroneous word segmentation results. In addition, broken characters are the main cause for erroneous character segmentation results.

Figure 10. Indicative text line segmentation errors.

Figure 11. Indicative word segmentation errors.

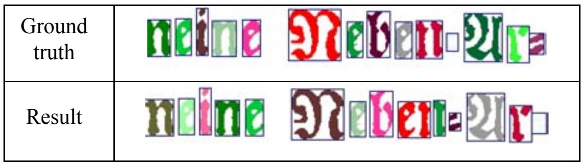


Figure 12. Indicative character segmentation errors.

In order to test the proposed evaluation scheme, we applied the FineReader Engine 8.1 [1] to the historical book and evaluated the final recognition result as well as the text line, word and character segmentation stages. Tables 2 and 3 summarize the evaluation results for all stages using FineReader Engine 8.1. The acceptance threshold is also set to $T_a = 95\%$ for text line segmentation and $T_a = 90\%$ for word and character segmentation. Although FineReader performs reasonably well for text line and word segmentation, we can observe that (i) character segmentation performance is quite low and (ii) character accuracy is remarkably low since the font of the historical book is not trained by the recognition engine. Table 4 lists the six most common errors ("confusions") made by the OCR system. Figure 13 shows the recognition result for a document image portion of the historical book as well as the corresponding ground truth.

Table 2. Evaluation results for text line, word and character segmentation using FineReader Engine 8.1.

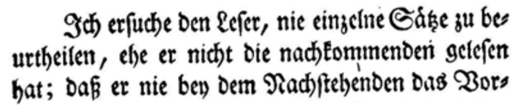
	N	M	o2o	DR	RA	FM
Text Lines	2647	2628	2593	97,9%	98,6%	**98,2%**
Words	18575	18854	17842	96,0%	94,6%	**95,2%**
Characters	14667	15961	12470	85,0%	78,1%	**81,4%**

Table 3. Recognition evaluation results using the FineReader Engine 8.1.

#characters	**116887**
#errors	**40494**
#insertions	2224
#deletions	32338
#substitutions	5932
Character Accuracy	**65,35%**

Table 4. The six most common errors ("confusions") made by the FineReader Engine 8.1.

#errors	Correct character	Generated character
3815	D	b
1720	S	f
984	Ch	d)
538	Ch	d;
484	K	f
477	Z	j


Ich ersuche den Leser, nie einzelne Sätze zu be-
urtheilen, ehe er nicht die nachkommenden gelesen
hat; daß er nie bey dem Nachstehenden das Vor-

(a)

Ich ersuche den Leser, nie einzelne Sätze zu be
urtheilen, ehe er nicht die nachkommenden gelesen
hat; daß er nie bey dem Nachstehenden das Vor

(b)

3ch erSuche den LcSer, nie einzelne Cade ju bc
urcheilen, ehe er nicht die nachkommenden gclefen
hat; da§ er nie bey dem Nachstehenden das Bor

(c)

Figure 13. Recognition result using FineReader Engine 8.1 (a) original document image; (b) correct document transcription; (c) recognition result.

4. CONCLUSIONS

A new comprehensive methodology has been proposed in order to evaluate the performance of noisy historical document recognition techniques. According to the proposed methodology, we evaluate not only the final noisy recognition result but also the main intermediate stages of text line, word and character segmentation. We mainly focus on the efficient creation of the text line, word and character segmentation ground truth guided by the transcription of the historical documents. In order to facilitate the text line segmentation ground truth creation, we use a Hough transform based methodology guided by the number of the text lines. Concerning the word and character segmentation ground truth, we use a gap classification technique constrained by the number of the words and characters for every text line. According to the experimental results, the percentage of time saved for the text line, word and character segmentation ground truth creation is more than 90%.

Acknowledgement

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement n° 215064 (project IMPACT).

5. REFERENCES

[1] ABBYY FineReader OCR: http://finereader.abbyy.com/

[2] Adamek, T., O'Connor, N.E. and Smeaton, A.F. 2007. Word Matching Using Single-Closed Contours for Indexing Handwritten Historical Documents. International Journal on Document Analysis and Recognition (IJDAR). Vol. 9, No 2-4 (2007), 153-165.

[3] Chang F., Chen C.-J. and Lu C.-J. A linear-time component-labeling algorithm using contour tracing technique. Comput. Vision Image Understanding. 93 (2) (2004), 206–220.

[4] Choudhury, G.S., DiLauro, T., Ferguson R., Droettboom, M. and Fujinaga I. 2006. Document Recognition for a Million Books. D-Lib Magazine. Vol. 12, No. 3 (2006).

[5] Gatos B., Antonacopoulos, A. and Stamatopoulos, N. 2007. Handwriting Segmentation Contest. 9th International Conference on Document Analysis and Recognition, Curitiba, Brazil, (September 2007), 1284-1288.

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

[7] Gezerlis, V.G. and Theodoridis, S. 2002. Optical Character Recognition for the Orthodox Hellenic Byzantine music notation. Pattern Recognition 35 (2002), 895-914.

[8] Kanai, J. Nartker, T.A. Rice, S. and Nagy, G. 1993. Performance metrics for document understanding systems. Proceedings of the Second International Conference on Document Analysis and Recognition, Tsukuba Science City, Japan, (1993), 424-427.

[9] Konidaris, T., Gatos, B., Ntzios, K., Pratikakis, I. And Perantonis S.J. 2007. Keyword – Guided Word Spotting in Historical Printed Documents Using Synthetic Data and User feedback. International Journal on Document Analysis and Recognition (IJDAR), Vol. 9, No 2 (2007) 167-177.

[10] Kornfield, E.M, Manmatha, R. and Allan J. 2004. Text Alignment with Handwritten Documents. In Proceedings of the 1st International Workshop on Document Image Analysis for Libraries (DIAL 2004), 195.

[11] Laskov, L. 2006. Classification and Recognition of Neume Note Notation in Historical Documents. In Proceedings of the International Conference of Computer Systems and Technologies (CompSysTech 2006), V.4.1-4.

[12] Lavrenko, V., Rath, T.M. and Manmatha R. 2004. Holistic Word Recognition for Handwritten Historical Documents. In Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL 2004), 278-287.

[13] Louloudis G., Gatos, B., Pratikakis, I. and Halatsis, C. 2008. Text line detection in handwritten documents, Pattern Recognition (41), No. 12, (December 2008), 3758-3772.

[14] Louloudis G., Gatos, B. and Pratikakis, I. 2008. Line and word segmentation of handwritten documents. In International Conference on Frontiers in Handwriting Recognition (ICFHR'08), Montreal, Canada, (August 2008), 247–252.

[15] Ntzios, K., Gatos, B., Pratikakis, I., Konidaris, T. and Perantonis S. J. 2007. An Old Greek Handwritten OCR System based on an Efficient Segmentation-free Approach. International Journal on Document Analysis and Recognition (IJDAR), Vol. 9, No. 2 (2007) 179-192.

[16] Phillips, I. and Chhabra, A. 1999. Empirical Performance Evaluation of Graphics Recognition Systems. IEEE Transactions Pattern Analysis and Machine Intelligence, Vol.21, No.9, (1999), 849-870.

[17] Rath, T.M. and Manmatha, R. 2007. Word Spotting for Historical Documents. International Journal on Document Analysis and Recognition (IJDAR), Vol. 9, No 2 (2007), 139-152.

[18] Stamatopoulos, N., Gatos, B. and Kesidis,A. 2007. Automatic Borders Detection of Camera Document Images. In 2nd Int. Workshop on Camera-Based Document Analysis and Recognition, Curitiba, Brazil, (2007), 71-78.

[19] Vamvakas, G., Gatos, B., Stamatopoulos, N., Louloudis G. and Perantonis, S.J. 2008. A Complete Optical Character Recognition Methodology for Historical Documents. In Proceedings of the DAS, (2008), 525-532.

[20] Wahl F.M., Wong K.Y. and Casey R.G. Block Segmentation and Text Extraction in Mixed Text/Image Documents. Computer Graphics and Image Processing, 20, (1982), 375-390.

[21] Zimmermann, M. and Bunke H. 2002. Automatic Segmentation of the IAM Off-line Database for Handwritten English Text. In Proceedings of the 16th International Conference on Pattern Recognition (ICPR 2002), 35-39.