Recognition and Enrichment of Archival Documents

D6.12.
Line and Word Segmentation Tools P3

Georgios Louloudis, Nikolaos Stamatopoulos, Basilis Gatos, NCSR Demokritos

Distribution:

http://read.transkribus.eu/

READ
H2020 Project 674943

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 674943

Project ref no.	H2020 674943
Project acronym	**READ**
Project full title	**Recognition and Enrichment of Archival Documents**
Instrument	H2020-EINFRA-2015-1
Thematic Priority	EINFRA-9-2015 - e-Infrastructures for virtual research environments (VRE)
Start date / duration	01 January 2016 / 42 Months

Distribution	Public
Contractual date of delivery	31/12/2018
Actual date of delivery	28/12/2018
Date of last update	30/11/2018
Deliverable number	D6.12
Deliverable title	Line and Word Segmentation Tools P3
Type	Demonstrator
Status & version	Public & version 1
Contributing WP(s)	WP6
Responsible beneficiary	NCSR
Other contributors	URO
Internal reviewers	UPVLC, EPFL
Author(s)	Georgios Louloudis, Nikolaos Stamatopoulos, Basilis Gatos
EC project officer	Martin Majek
Keywords	Text Line Segmentation, Word Segmentation

Table of Contents

Executive Summary

This deliverable reports on the achievements concerning the tasks of text line and word segmentation at the end of the third year of the READ project. The deliverable consists of two parts. The first part (Text Line Segmentation) contains a brief description of the work accomplished for the task of text line segmentation. The excellent method proposed by the Rostock partner (CITlab LA module) during the second year of the project which was the winning method of the ICDAR 2017 competition on baseline detection (cBAD) [Diem2017] is the one currently included in the Transkribus platform. The great success of the Rostock method is not only related with the method's excellent accuracy but also with the fact that no prior segmentation of the document image into text regions is necessary. The astonishing performance of the Rostock method is the reason to shift resources to other tasks of the segmentation pipeline such as Layout Analysis and Word Segmentation. In this deliverable, we include a short description of the different options that can be selected by the user of the system for the text line segmentation method on the Transkribus platform. The second part of this deliverable (Word Segmentation) describes a novel method developed by the NCSR group for the segmentation of a document image into words without the need of any prior segmentation step (as the method that was developed for the text line segmentation task at the first part). The method is applied directly to the initial grayscale image, producing several word segmentation hypotheses which can then be used by the Query by Example keyword spotting method in order to produce the final ranking list of similar words. The experiments conducted on three different datasets by comparing several scenarios (existence of segmentation results on different levels) prove the superiority of the proposed word segmentation method. Finally, it should be noted that the NCSR word segmentation method is planned to be submitted to the next ICDAR conference (ICDAR 2019).

1. Text Line Segmentation

One of the early tasks in a handwriting recognition system is the segmentation of a handwritten document image into text lines, which is defined as the process of defining the region of every text line on a document image. Several challenges exist on historical documents which should be addressed by a text line segmentation method. These challenges include: a) the difference in the skew angle between lines on the page or even along the same text line, b) overlapping and touching text lines, c) additions above the text line and d) deleted text. Figure 1.1 presents one example for each of these challenges.

Two main variations exist for representing the results of a text line segmentation method: i) using polygons that enclose the corresponding text lines and ii) using baselines i.e. a set of (poly)line segments which correspond to the imaginary lines on which the scribe writes the text. Figure 1.2 presents one example of each of the abovementioned representation variations.

As it was mentioned on the second year's report (D.6.11) since the baseline representation has the advantage of needing less time for correction and since according to [Romero2015] the baseline representation produces comparable results in terms of HTR accuracy with the polygon representation, it was decided to use the baseline representation for the description of the text line segmentation results.

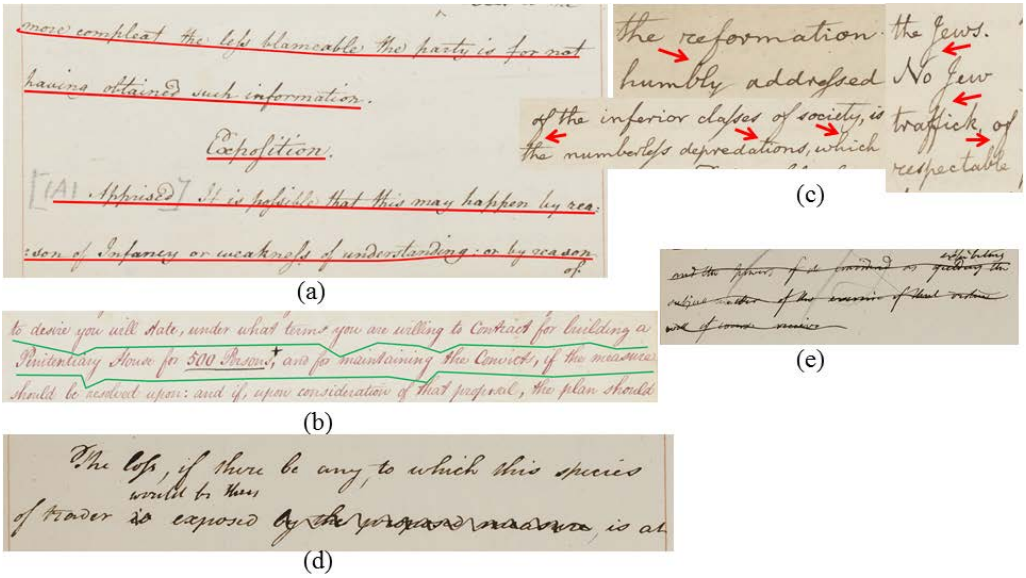


Figure 1.1: Challenges encountered on historical document images for text line segmentation: (a) Difference in the skew angle between lines on the page or even along the same text line, (b) overlapping text lines, (c) touching text lines, (d) additions above a text line, e) deleted text.

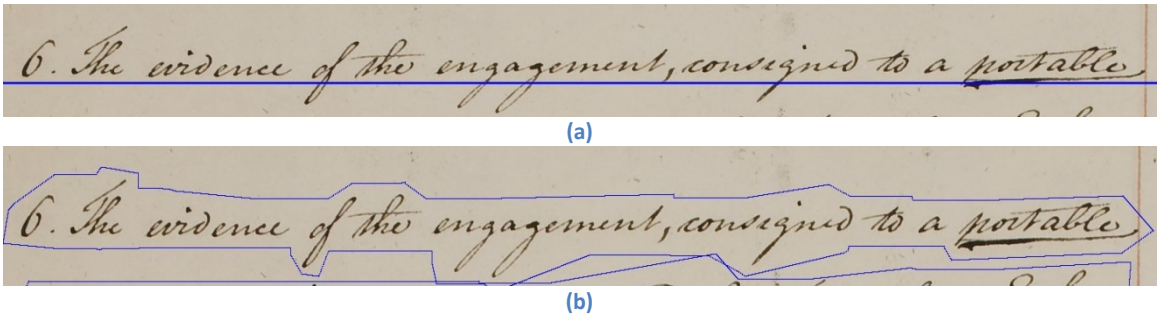


Figure 1.2: Representation of the text line segmentation result using (a) baseline and (b) polygon.

As described in [Gruning2018], the CITlab (advanced) LA module relies on a deep neural network. A default network which was trained on the cBAD train set (leading to F-values of 97.8 and 91.6 on the cBAD simple and complex test sets, respectively) is chosen with the "Preset" option (see Figure 1.3). By choosing the "Text orientation: Default" setting it is assumed that the entire text in the image is 0° (+-10°) oriented. The "Homogeneous" setting allows for orientations of 0°, 90°, 180° and 270°, but homogeneously. I.e., all text lines are forced to have the SAME orientation. To allow for the detection of various orientations, the "Preset" neural network was trained on arbitrarily (modulo 90°) oriented images of the cBAD training set. Finally, the "Heterogeneous" option allows for mixed orientations, e.g., 0° text lines along with a 90° oriented text lines could be detected. However, since the task to detect oriented text lines is harder than the task to detect 0° text lines and therefore the results in the 0° case are usually slightly better, it is recommended to choose the "Default" setting for collections which contain mainly text in the 0° orientation, which is true for most of the historical collections.

As it was described in [Gruning2018], the neural network not only detects baselines but also separators. These separators are used to detect the structure of the text, e.g., the beginning and end of text lines in a table. However, if text regions are already available it could be counterproductive to detect separators within these regions. Therefore, the "Use

separators: Default" setting utilizes the detected separators solely if no text regions are given. The "Never" and "Always" options are self-explanatory.

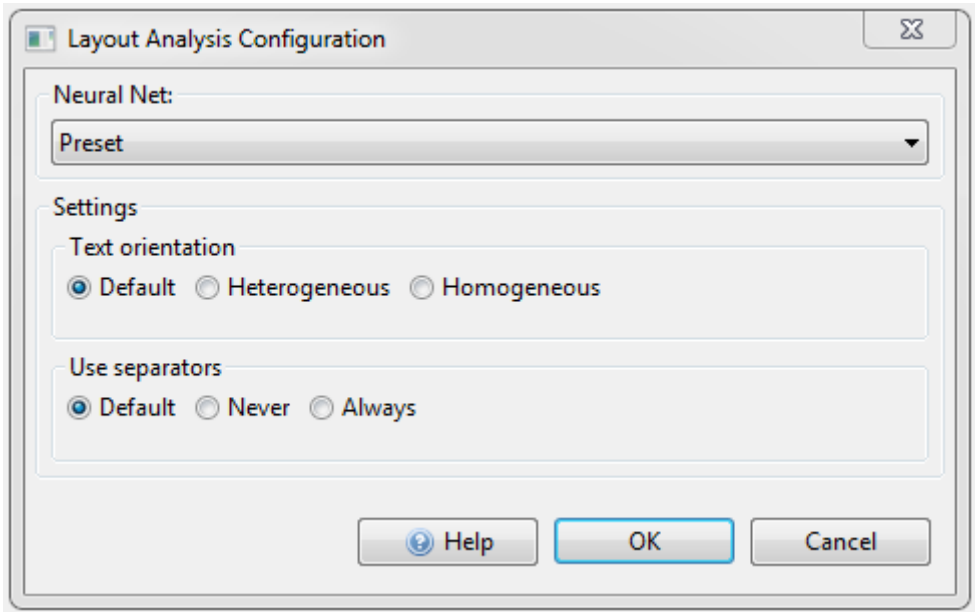


Figure 1.3: Screenshot of the text line segmentation configuration menu on the Trankribus platform.

2. Word Segmentation

Word segmentation refers to the process of defining the word regions of a text line. Since nowadays most handwriting recognition methods assume text lines as input, the word segmentation process is usually necessary only for segmentation-based query by example (QbE) keyword spotting (KWS) methods. Segmentation of historical handwritten document images still presents significant challenges and it is an open problem. These challenges include the appearance of skew along a single text line, the existence of slant, the non-uniform spacing of words as well as the existence of punctuation marks (Figure 2.1).

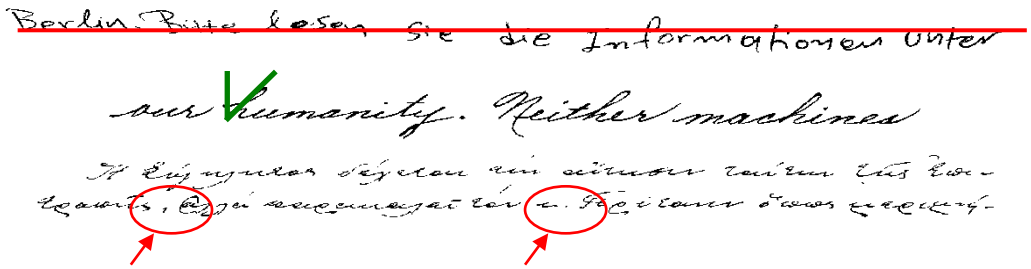
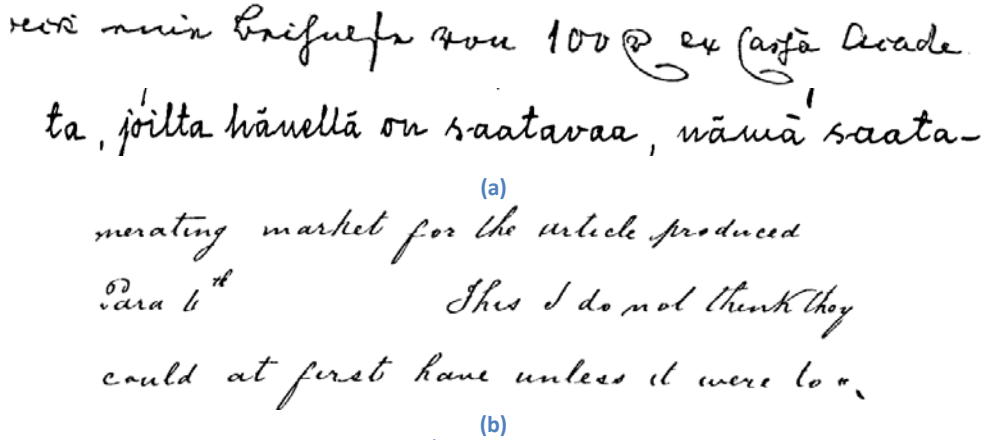


Figure 2.1: Challenges encountered on historical document images for word segmentation.

2.1. NCSR Word Segmentation Method – 3rd Year

In the frame of “READ” project a word segmentation method (NCSR 2nd Year) was delivered in the second year. This method was an extension of the method presented in [Louloudis2009], adapted to historical handwritten documents in order to cope with common challenges such as the existence of long ascenders/descenders (Figure 2.1.1(a)) and the presence of extreme values/outliers (e.g. large distances of adjacent words) (Figure 2.1.1(b)).

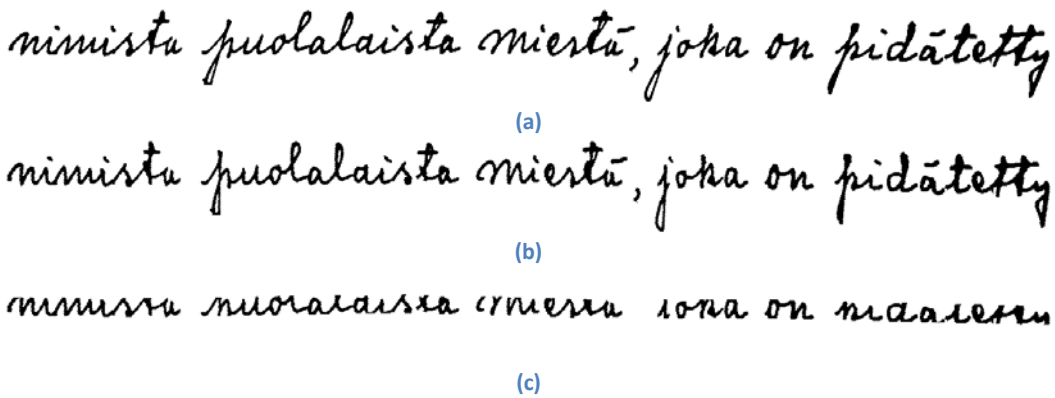


(a)

(b)

Figure 2.1.1: Challenges addressed by the NCSR 2nd Year method: (a) existence of long ascenders/descenders; (b) presence of extreme values/outliers (i.e. large distances of adjacent words).

NCSR word segmentation method contains two steps. The first step deals with the computation of the Euclidean distances of adjacent components using only the main zone of the text line image in order to exclude the ascenders/descenders as well as the punctuation marks (see Figure 2.1.2).



(a)

(b)

(c)

Figure 2.1.2: (a) Original text line image; (b) after slant correction; (c) after main zone detection.

The second step concerns the classification of the previously computed distances as either inter-word gaps or intra-words distances using the Student's-t distribution. The main advantage of the Student's-t distribution concerns its robustness to the existence of outliers.

It should be stressed that the NCSR word segmentation method is developed in C++ following the guidelines of the Transkribus interface and it is available at github:

https://github.com/Transkribus/NCSR_Tools

In the third year of the "READ" project, a novel word segmentation method (NCSR 3rd Year) was developed providing multiple hypothesis segmentation results (Figure 2.1.3) since the word segmentation process is mainly used as part of a segmentation-based Query-by-Example (QbE) keyword spotting method. The main advantage of the new segmentation method is that it can be applied directly to the grayscale document image. As a result, preprocessing methods such as binarization and segmentation steps (e.g. layout or text line) are unnecessary thus reducing the corresponding processing time.

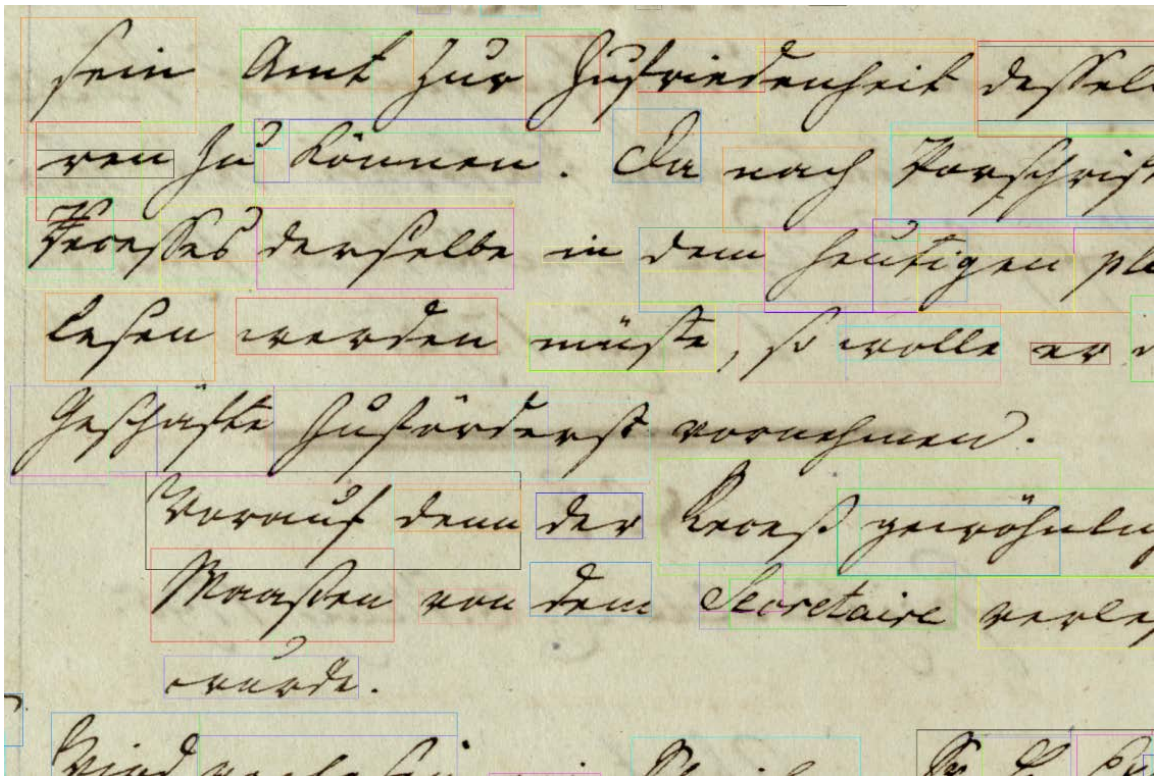


Figure 2.1.3: An example of a multiple hypothesis word segmentation result.

The new word segmentation method is based on the MSER technique (Maximally Stable Extremal Regions) [Donoser2006]. MSER is a method for blob detection in images. The MSER algorithm extracts a number of co-variant regions from an image, called MSERs: an MSER is a stable connected component of some gray-level sets of the image. It is based on the idea of taking regions which stay nearly the same through a wide range of thresholds. The word extremal refers to the property that all pixels inside the MSER have either higher (bright extremal regions) or lower (dark extremal regions) intensity than all the pixels on its outer boundary (Figure 2.1.4). Once the initial MSER regions have been detected (Figure 2.1.5(a)) a post processing step is applied in order to detect possible word regions by combining MSER regions (Figure 2.1.5(b)). Our goal is to increase the number of correctly segmented words (high Recall) while at the same time retain the total number of detected words low (high Precision).

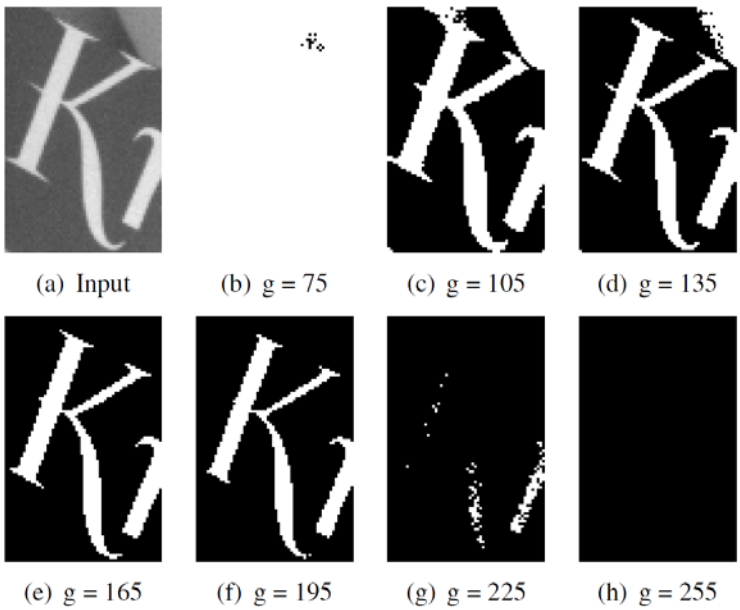
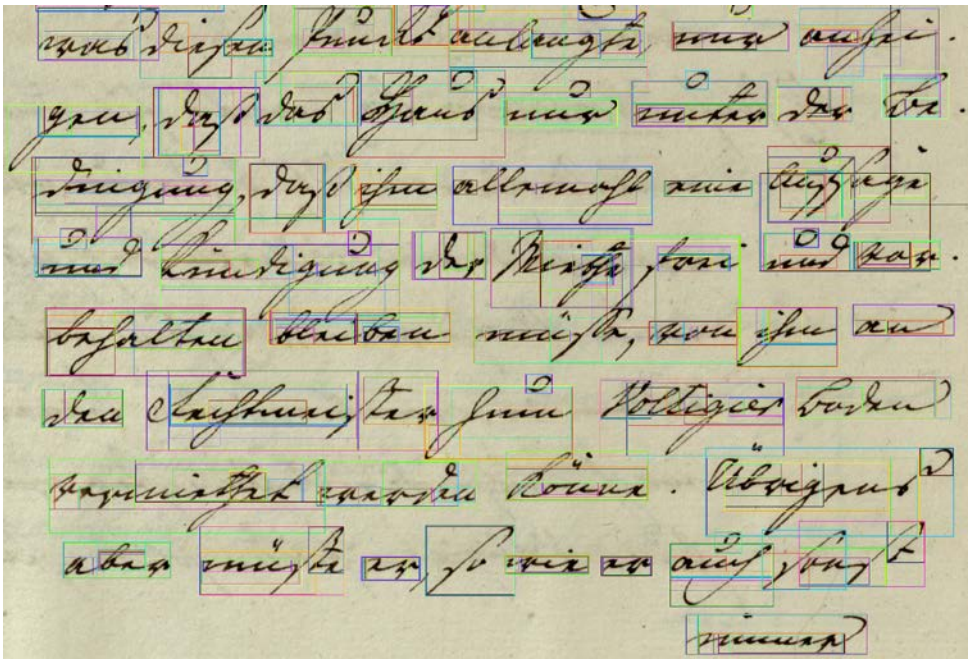


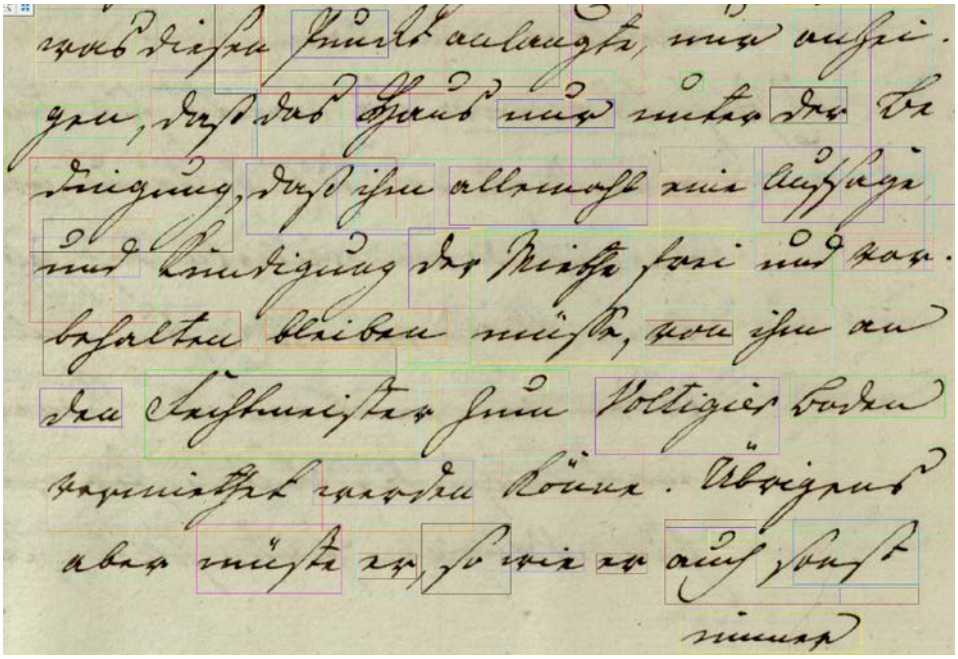
Figure 2.1.4: MSRE example on a document image: Threshold images analyzed during creation of component tree. Figure (a) shows the considered area and figures (b) to (g) the results of thresholding this image at gray level g. The letter k is identified as MSER because the size of the connected region does not change significantly in the gray level range from 135 to 195. [Donoser2006]

2.2. Evaluation

We indirectly evaluated the performance of the new word segmentation method by measuring the segmentation-based QbE KWS performance starting from several word segmentation results. These results were produced by making use of a variety of assumptions concerning the segmentation of the document image into specific entities. Table 2.2.2 presents the performance of the NCSR-POG (Projections of Oriented Gradients) method [Retsinas2016] using several segmentation scenarios in order to produce the words. Starting from the first row of Table 2.2.2, it is assumed that all segmented entities (text regions referred as layout, text lines and words) correspond to ground truth entities. The second row of Table 2.2.2 assumes that regions and text lines correspond to ground truth entities whereas the words were produced by an automatic method (NCSR 2nd Year Word Segmentation). Going to the third row, in addition to the automatic word segmentation method, the text lines were also produced automatically using the NCSR 2nd Year Text Line Segmentation method. The fourth row assumes that the whole pipeline was produced by involving an automatic method to all intermediate steps (i.e. layout analysis, text line and word segmentation). As it is expected, as we go from the first row to the fourth row of Table 2.2.2, the performance of the keyword spotting method decreases since more automatic steps are involved for the production of the word segmentation result. The final row corresponds to the application of the novel word segmentation method described in this deliverable. It is evident from the table that this method is applied directly to the initial grayscale image. It should be noted that for all scenarios where a binary image was necessary, we used the NCSR 1st Year binarization method.



(a)



(b)

Figure 2.1.5: NCSR multiple hypothesis segmentation method: (a) Initial MSRE regions (b) final word hypothesis.

The performance of the word spotting methods was recorded in terms of the Mean Average Precision (MAP) on three challenging datasets of historical handwritten documents: (i) Konzilsprotokolle (GE), (ii) NAF (FN) and (iii) BL (EN). Table 2.2.1 summarizes the number of documents as well as the number of words for each dataset. Time and memory requirements are recorded in terms of the following metrics which are self-explanatory: Retrieval Time per Query (RTpQ) and Size per Document (SpD). For more details for the datasets, the KWS method and the evaluation protocol see also Deliverable D7.13 “Keyword Spotting Engines: QbE, QbS P1”.

Table 2.2.1: Summary of dataset information used to evaluate the word segmentation method.

Dataset	#documents	#text lines	#words
Konzilsprotokolle (GE)	100	2555	15567
NAF (FN)	56 (double pages)	3186	16201
BL (EN)	115	2971	15739

As the experimental results indicate, the NCSR-POG method achieves the best performance using the words (multiple hypothesis) produced by the new NCSR 3rd year method. We should mention that there are datasets, i.e. GE and EN, in which the NCSR-POG method achieves better results using the multiple hypothesis segmentation results even for the case starting from the ground-truth words. Table 2.2.3 presents time and memory requirements of the KWS pipeline using different segmentation frameworks. As we can see, the requirements with respect to time and memory of the NCSR-POG method are higher using the new NCSR 3rd year method since more possible words are produced. However, the total processing time of a document is significant lower using the new word segmentation method since all the other preprocessing and segmentation steps have been removed.

Table 2.2.2: Comparative experimental results of NCSR-POG method using several segmentation scenarios

Binarization	Layout	Text Lines	Words	GE MAP (%)	EN MAP (%)	FN MAP (%)
-	GT	GT	GT	58.15	42.20	65.34
NCSR 1st Year	GT	GT	NCSR 2nd Year	54.45	37.27	62.82
NCSR 1st Year	GT	NCSR 2nd Year	NCSR 2nd Year	53.67	35.97	62.59
NCSR 1st Year	NCSR 1st Year	NCSR 2nd Year	NCSR 2nd Year	50.66	33.44	60.17
-	-	-	NCSR 3rd Year	**59.40**	**42.73**	**63.20**

Table 2.2.3: Time and Memory Requirements for the NCSR-POG method using several segmentation scenarios

KWS pipeline	Preprocessing & Segmentation (sec/doc)	RTpQ(sec)	SpD(KB)
NCSR-POG using NCSR 2nd Year segmentation method	25	0.0076	97
NCSR-POG using NCSR 3rd Year Word segmentation method	2	0.0228	291

3. References

[Donoser2006] M. Donoser and H. Bischof “Efficient Maximally Stable Extremal Region (MSER) Tracking” Conference on Computer Vision and Pattern Recognition (CVPR’06), pp. 553-560, 2006.

[Louloudis2009] G. Louloudis, B. Gatos, I. Pratikakis, C. Halatsis, "Text line and word segmentation of handwritten documents", Pattern Recognition, vol. 42, no 12, pp. 3169-3183, 2009.

[Romero2015] V. Romero, J.A. Sanchez, V. Bosch, K. Depuydt, and J. de Does, “Influence of text line segmentation in handwritten text recognition”, 13th International Conference on Document Analysis and Recognition, pp. 536-540, 2015.

[Diem2017] Markus Diem, Florian Kleber, Stefan Fiel, Tobias Grüning, and Basilis Gatos: “cBAD: ICDAR2017 Competition on Baseline Detection”, in Proceedings of the 14th International Conference on Document Analysis and Recognition (2017), 1355-1360.

[Gruning2018] Tobias Grüning, Gundram Leifert, Tobias Strauß and Roger Labahn: “A Two-Stage Method for Text Line Detection in Historical Documents”, arXiv preprint arXiv:1802.03345.

[Retsinas2016] G. Retsinas, G. Louloudis, N. Stamatopoulos and B. Gatos, “Keyword Spotting in Handwritten Documents using Projections of Oriented Gradients”, 12th Workshop on Document Analysis Systems (DAS'16), pp. 411-416, Santorini, Greece, 2016.