Efficient Transcript Mapping to Ease the Creation of Document Image Segmentation Ground Truth with Text-Image Alignment

Nikolaos Stamatopoulos, Georgios Louloudis, Basilis Gatos
Computational Intelligence Laboratory, Institute of Informatics and Telecommunications
National Center for Scientific Research "Demokritos"
GR-153 10 Athens, Greece
{nstam, louloud, bgat}@iit.demokritos.gr

***Abstract*— One of the major issues in document image processing is the efficient creation of ground truth in order to be used for training and evaluation purposes. Since a large number of tools have to be trained and evaluated in realistic circumstances, we need to have a quick and low cost way to create the corresponding ground truth. Moreover, the specific need for having the correct text correlated with the corresponding image area in text line and word level makes the process of ground truth creation a difficult, tedious and costly task. In this paper, we introduce an efficient transcript mapping technique to ease the construction of document image segmentation ground truth that includes text-image alignment. The proposed text line transcript mapping technique is based on Hough transform that is guided by the number of the text lines. Concerning the word segmentation ground truth, a gap classification technique constrained by the number of the words is used. Experimental results prove that using the proposed technique for handwritten documents, the percentage of time saved for ground truth creation and text-image alignment is more than 90%.**

Keywords-ground truth creation; transcript mapping; document image segmentation

I. INTRODUCTION

Efficient ground truth creation is essential for training and evaluation purposes in the document image analysis and recognition pipeline. Since a large number of tools have to be trained and evaluated in realistic circumstances we need to have a quick and low cost way to create the corresponding ground truth. Moreover, the specific need for having the correct text correlated with the corresponding image area in text line and word level (see Fig.1) makes the process of ground truth creation a difficult, tedious and costly task. 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. They are mainly based on hidden Markov models (HMMs) [1-3] and dynamic time warping (DTW) [4-6] and mainly focus on the alignment of handwritten document images with the corresponding transcription on word level.

In this paper, we introduce an efficient transcript mapping technique to ease the construction of document image segmentation ground truth that includes text-image alignment in text line and word level. We facilitate the annotation of text line and word segmentation ground truth regions as well as the correlation with corresponding text making use of the correct document transcription. In the proposed framework, we assume that the transcription includes the correct text line break information. This information is used in a novel transcript mapping module in order to efficiently create the text line and word segmentation ground truth. The proposed text line transcript mapping technique is based on Hough transform that is guided by the number of the text lines in order to efficiently create the text line segmentation result. Concerning the word segmentation ground truth, a gap classification technique constrained by the number of the words is used. The flowchart of the proposed methodology for the creation of document image segmentation ground truth that includes text-image alignment is demonstrated in Fig. 2. We recorded that using the proposed technique for handwritten documents, the percentage of time saved for ground truth creation and text-image alignment is more than 90%.

The remainder of the paper is organized as follows. In Section II, related work is presented. In Section III, the proposed methodology is detailed while experimental results are discussed in Section IV. Finally, conclusions are drawn in Section V.

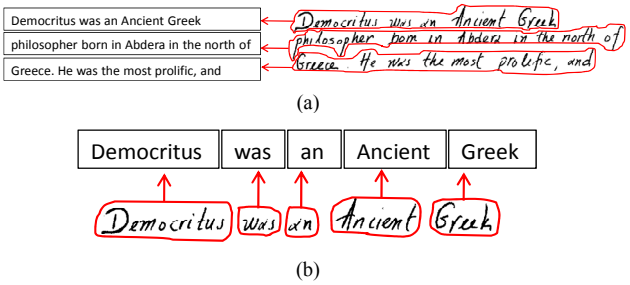

Figure 1. Correlation of the transcription with the corresponding image area: (a) text line level; (b) word level.

II. RELATED WORK

Transcript mapping techniques can be classified into two main categories according to the algorithm that is used for the alignment. The first category contains transcript mapping techniques that are based on hidden Markov models (HMMs). Zimmermann and Bunke [1] present an automatic segmentation scheme for cursive handwritten text lines using

978-0-7695-4221-8/10 $26.00 © 2010 IEEE
DOI 10.1109/ICFHR.2010.43

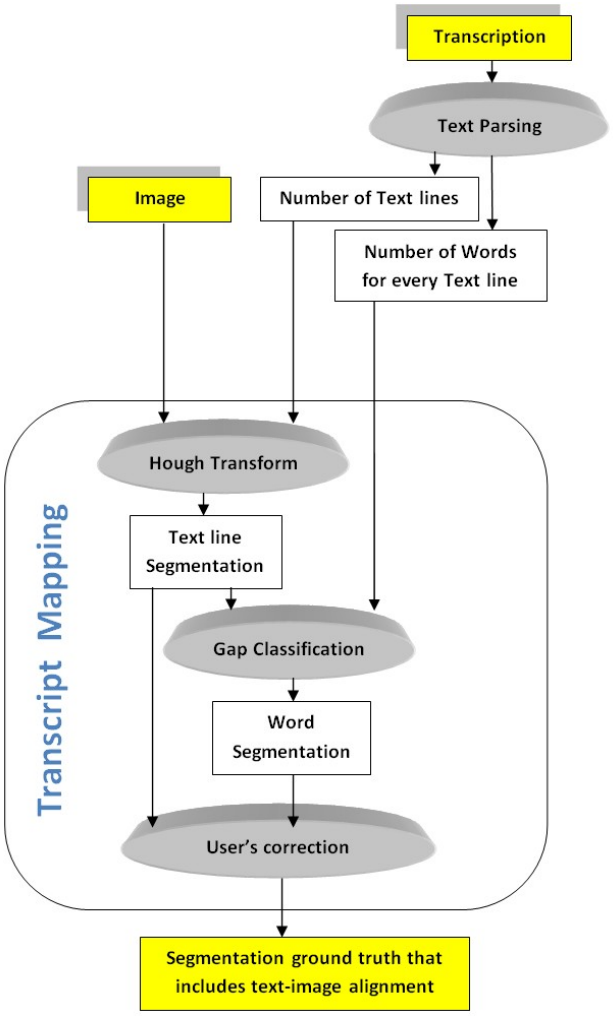

Figure 2. The flowchart of the proposed methodology for the creation of document image segmentation ground truth that includes text-image alignment.

transcriptions of the text lines and a HMM *based* recognition system. The segmentation scheme consists of two steps. In the first step the Viterbi decoder is used in forced alignment mode using a normalized text line, in which optimal word boundaries can be computed. The second step uses these word boundaries to assign the connected components of the normalized line to the individual words. The authors report a 98% word segmentation rate on the IAM database. In this same direction, Toselli et al. [2] propose an alignment method based on the Viterbi algorithm to find mappings between word images of a given handwritten document and their respective words on its transcription. This method takes advantage of the implicit alignment made by Viterbi decoding used in text recognition with HMMs. In [3], Rothfeder et al. use a linear HMM to solve the alignment problem without performing word recognition explicitly for each word image. All the word images are treated as the hidden variables, while the feature vectors are extracted from each of the word images, being modeled as observed variables. The Viterbi algorithm was used to decode the sequence of assignments to each of the word images. They evaluate the method on a set of 70 pages of George Washington collection and an average accuracy of 72.8% was reported.

The second category contains transcript mapping techniques that are based on dynamic programming or dynamic time warping (DTW) which is an algorithm for aligning two time series by minimizing the distance between them. Kornfield et al. [4] rely on DTW using as series the image locations from the segmentation step and the text words in transcription. This method does not require to perform word recognition for each segmented word image and it was applied to the historical handwritten documents of Washington collection and achieved 60.5% accuracy when aligning full pages at time. Lorigo and Govindaraju [5] propose a transcript mapping method for handwritten Arabic documents. It is based on an extension of DTW that uses true distances when mapping multiple entries from one series to a single entry in the second series. In [6], Jawahar and Kumar present a hierarchical approach for transcript mapping of printed Indian documents on character level which is based on DTW. They use automatic and semi-automatic annotation tools for text line and word alignment as well as several validation tools. Tomai et al. [7] propose a method in order to limit the lexicon of a handwriting recognizer by using the transcription. A ranked list of possible words from the lexicon is returned for each recognized word image. Several different likely segmentations of a line are made. Then, word mapping is defined using word recognition results by a dynamic programming algorithm that finds the best match. If a mapping cannot be performed with enough high confidence for a word then it is omitted. Huang and Srihari [8] present a recognition-based alignment algorithm. A word recognizer generates multiple choices as a result. Then, a dynamic programming is used to find the optimal alignment between two words strings: the first one is the truth from the transcription and the second one is the recognition results from the word recognition. The authors report 84.7% accuracy in aligning words on 20 pages from a handwritten database. In [9], Zinger et al. present a method on text-image alignment in context of building a historical document retrieval system. The images of handwritten lines are automatically segmented from the scanned pages of historical documents and then manually transcribed. Alignment on word level is based on the longest spaces between portions of handwriting is a baseline. To take into account the relative word length, the expressions for the cost function that has to be minimized for aligning text words with their images is defined. Finally, a different technique is presented in [10] by Hobby that assumes that along with the transcription there is a page description that denotes where the words in the transcription appear on the page. It tries to find a geometric transformation between the document description and the image of the document which minimizes a cost function. An overview of the characteristics of existing transcript mapping techniques is presented in Table I.

TABLE I. AN OVERVIEW OF TRANSCRIPT MAPPING TECHNIQUES.

Transcript mapping technique	Year	Alignment Method	Alignment Level	Document Images
Zimmermann and Bunke [1]	2002	HMM	Text Line Word	Handwritten
Toselli et al. [2]	2007	HMM	Word	Historical Handwritten
Rothfeder et al. [3]	2006	HMM	Word	Historical Handwritten
Kornfield et al. [4]	2004	DTW	Word	Historical Handwritten
Lorigo and Govindaraju [5]	2007	DTW	Word	Handwritten Arabic
Jawahar and Kumar [6]	2007	DTW	Text Line Word Character	Printed Indian
Tomai et al. [7]	2002	Dynamic Programming	Word	Historical Handwritten
Huang and Srihari [8]	2006	Dynamic Programming	Word	Handwritten
Zinger et al. [9]	2009	Cost function	Word	Historical Handwritten
Hobby [10]	1998	Geometric transformation	Character	Printed

III. PROPOSED METHODOLOGY

According to the proposed methodology, in order to ease the construction of document image segmentation ground truth that includes text-image alignment we follow several distinct steps. 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 for every text line. This information is used in a transcript mapping module in order to efficiently create the text line and word segmentation ground truth. 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 segmentation ground truth, we first use a gap classification technique constrained by the number of the words for every text line indicated at the text parsing stage. For the creation of the final text line and word segmentation ground truth we also involve a user guided correction module. As it will be demonstrated in Section IV, only a small number of segmentation results needs correction since the proposed automatic transcript mapping technique has been proved efficient and time saving. All involved steps are detailed in this section.

A. Text Parsing

The transcription contains useful information that can be used in order to correctly segment a document image into text lines and words. 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 of every text line i .

B. Transcript Mapping

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

1) Text line segmentation: The methodology for the segmentation of a document image into text lines is a modification of the methodology described in [11] which takes into consideration the number of text lines NL calculated at the Text Parsing phase (Section *III-A*). It includes two stages: (a) Hough transform mapping and (b) post-processing.

a) Hough Transform Mapping: At first, the connected components [12] (CCs) of the document image are extracted and then, the average character height AH of the whole document image is calculated based on the average height of all CCs.

The main stage of the text line segmentation methodology is the application of the Hough transform on a set of points (see [11]). 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:

$$x\cos(\theta) + y\sin(\theta) = p \quad (1)$$

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 $(p_i-5,\theta_i)...(p_i+5,\theta_i)$. In [11], the cell (p_i,θ_i) having the maximum contribution is calculated until a stopping criterion is met. In our approach, the stopping criterion is that the number of maximum cells (p_i,θ_i) should not exceed the number of text lines NL calculated at the Text Parsing phase (Section III-A).

b) 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. Text line segmentation result of a handwritten document image portion is presented in Fig. 3.

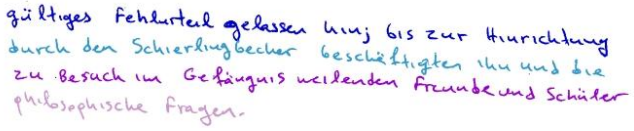

Figure 3. Text line segmentation result of a handwritten document image portion.

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 concerns 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 [13] that is guided by the expected number of words NW_i.

a) 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.

b) 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 *III-A*) as follows:

Let L be the number of the overlapped components of the text line image. The total number of distances computed is $L-1$. 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. In Fig. 4 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 Fig. 4(d). Word segmentation result of a handwritten document image portion is presented in Fig. 5.

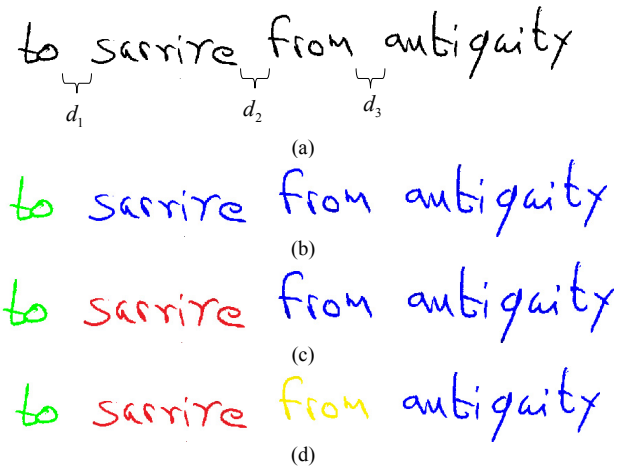

Figure 4. An example of word segmentation results using different segmentation thresholds; (a) initial image with the first three distances after the sorting; word segmentation results generated after considering as threshold the distance (b) d1; (c) d2; and (d) d3 Threshold d3 yields the correct result, since the expected number of words NW_i is 4.

Figure 5. Word segmentation result of a handwritten document image portion.

C. Correction of Segmentation Resutls

Once text lines and words have been detected, making use of the document transcription (Section III-B), 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. 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 IV, only a small number of segmentation results needs correction since the proposed automatic transcript mapping technique has been proved efficient and time saving.

IV. Experimental Results

The proposed transcript mapping methodology was tested on the set that used for the ICDAR2009 Handwriting Segmentation Contest [14] and consists of 200 document images that contain 4034 text lines and 29717 words. First, we measured the performance of the proposed technique to detect the ground truth regions before the user's intervention and we compared it with the performance of the participating methods as well as with two state-of-the-art techniques for document image segmentation. We compared with RLSA (implementation based on [15]) and Projection Profiles (implementation based on [16]) as well as with the 5 top ranked segmentation algorithms of the ICDAR2009 Handwriting Segmentation Contest (ILSP-LWSeg-09, PAIS, CMM, CUBS and CASIA-MSTSeg) [14]. The segmentation performance was measured in terms of Detection Rate (DR), Recognition Accuracy (RA) and F-Measure (FM) (as in [14]) while the acceptance threshold to have a one-to-one match (o2o) was set to 95% for text line segmentation and 90% for word segmentation.

We measured the total time needed for the creation of the segmentation ground truth that includes text-image alignment for the cases of (i) completely manual creation, (ii) first applying a state-of-the-art segmentation methodology and then correction, and (iii) applying the proposed transcript mapping methodology and then correction. We recorded that for the manual creation of the text line and word ground truth for one document image the user needs 280 and 600 seconds in average, respectively. We also estimated that the average time needed for the user to correct a word segmentation error was 5 seconds while to correct a text line segmentation error was 10 seconds using the tool described in Section III-C. We also took into account an average time for visually checking the generated text line and word segmentation result which was 13 and 27 seconds per image, respectively. We also assumed that if we have a completely corrected segmentation result, the correspondence with the text is done automatically.

All results are presented in Tables II and ranked according to the estimated time that is needed to create a segmentation ground truth that includes text-image alignment. As it can be observed, by using the proposed methodology, the percentage of time saved for ground truth creation compared to the complete manual process is about 92%. If we compare with the process of first applying the best working state-of-the-art segmentation algorithm and then applying correction, the percentage of time saved is about 12%.

V. Conclusions

In this paper, we introduce an efficient transcript mapping technique to ease the construction of document image segmentation ground truth that includes text-image alignment. The proposed text line transcript mapping technique is based on Hough transform that is guided by the number of the text lines while for word segmentation ground truth, a gap classification technique constrained by the number of the words is used. Using the proposed technique for handwritten documents, the percentage of time saved for the text line and word segmentation ground truth creation is more than 90% compared to the manual process and 12% compared to the process of first applying the best working state-of-the-art segmentation algorithm.

Acknowledgment

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

TABLE II. COMPARATIVE RESULTS USING HANDWRITTEN SET.

	Text Line Detection				Word Detection				Estimated Time needed to create Segmentation Ground Truth that includes Text-Image Alignment
	o2o	DR (%)	RA (%)	FM (%)	o2o	DR (%)	RA (%)	FM (%)	
Manual Creation	-	-	-	-	-	-	-	-	**48.9 hours**
Projection Profiles	2538	62.9	57.8	60.2	20143	67.8	52.5	59.2	**19.7 hours**
RLSA	1615	40.0	38.9	39.4	24006	80.8	77.7	79.2	**16.9 hours**
CASIA-MSTSeg	3867	95.9	95.5	95.7	25938	87.3	82.6	84.8	**7.9 hours**
CUBS	4016	99.6	99.5	99.5	26631	89.6	84.4	87.0	**6.6 hours**
CMM	3975	98.5	98.3	98.4	27078	91.1	86.8	88.9	**6.0 hours**
PAIS	3973	98.5	98.6	98.5	27288	91.8	89.3	90.5	**5.8 hours**
ILSP-LWSeg-09	4000	99.2	98.9	99.0	28279	95.1	94.4	94.8	**4.3 hours**
Proposed Methodology	3913	97.0	97.1	97.1	28845	97.1	97.2	97.1	**3.8 hours**

REFERENCES

[1] M. Zimmermann and H. Bunke, "Automatic Segmentation of the IAM Off-line Database for Handwritten English Text" Int. Conference on Pattern Recognition, 2002, pp. 35-39.

[2] A. Toselli, V. Romero, E. Vidal, "Viterbi based alignment between text images and their transcripts" Workshop on Language Technology for Cultural Heritage Data, 2007, pp.9-16.

[3] J. Rothfeder, R. Manmatha and T.M. Rath, "Aligning Transcripts to Automatically Segmented Handwritten Manuscripts", Workshop on Document Analysis Systems, (DAS), 2006, pp. 84-95.

[4] E.M. Kornfield, R. Manmatha and J. Allan, "Text Alignment with Handwritten Documents" Int. Workshop on Document Image Analysis for Libraries, 2004, pp. 195-211.

[5] L. Lorigo and V. Govindaraju, "Transcript Mapping for Handwritten Arabic Documents", 14th SPIE Conference on Document Recognition and Retrieval, 2007, vol. 6500.

[6] C.V. Jawahar, A. Kumar, "Content-level Annotation of Large Collection of Printed Document Images" Int. Conference on Document Analysis and Recognition, 2007, pp.799-803.

[7] C. Tomai, B. Zhang and V. Govindaraju, "Transcript Mapping for Historic Handwritten Documents", International Workshop on Handwriting Recognition, 2002, pp. 413-418.

[8] C. Huang and S.N. Srihari, "Mapping transcripts to handwritten text" 10th International Workshop on Frontiers in Handwriting Recognition, 2006, pp. 15-20.

[9] S. Zinger, J. Nerbonne, and L. Schomaker, "Text-image alignment for historical handwritten documents", Proceedings of SPIE - The International Society for Optical Engineering, Document Recognition and Retrieval XVI, 2009, vol. 7247, pp. 1-8.

[10] J.D. Hobby, "Matching document images with ground truth", International Journal on Document Analysis and Recognition, vol. 1, no. 1, 1998, pp. 52-61.

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

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

[13] G. Louloudis, B. Gatos and I. Pratikakis, "Line and word segmentation of handwritten documents" Int. Conference on Frontiers in Handwriting Recognition, 2008, pp. 247–252.

[14] B. Gatos, N. Stamatopoulos, G. Louloudis, "ICDAR2009 Handwriting Segmentation Contest", Int. Conference on Document Analysis and Recognition, 2009, pp. 1393-1397.

[15] T. Konidaris, B. Gatos, K. Ntzios, I. Pratikakis, S. Theodoridis and S.J. Perantonis, "Keyword-guided word spotting in historical printed documents using synthetic data and user feedback", International Journal on Document Analysis and Recognition (IJDAR) 9 (2–4), 2007, pp. 167–177.

[16] A. Antonacopoulos, D. Karatzas, "Semantics-based content extraction in typewritten historical documents", Int. Conference on Document Analysis and Recognition, 2005, pp. 48–53.G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*