Text Line Detection for Heterogeneous Documents

Markus Diem, Florian Kleber and Robert Sablatnig
Computer Vision Lab
Vienna University of Technology
Email: diem@caa.tuwien.ac.at

***Abstract*—Text line detection is a pre-processing step for automated document analysis such as word spotting or OCR. It is additionally used for document structure analysis or layout analysis. Considering mixed layouts, degraded documents and handwritten documents, text line detection is still challenging. We present a novel approach that targets torn documents having varying layouts and writing. The proposed method is a bottom up approach that fuses words, to globally minimize their fusing distance. In order to improve processing time and further layout analysis, text lines are represented by oriented rectangles. Even though, the method was designed for modern handwritten and printed documents, tests on medieval manuscripts give promising results. Additionally, the text line detection was evaluated on the ICDAR 2009 and ICFHR 2010 Handwriting Segmentation Contest datasets.**

I. Introduction

Text line detection is an ongoing research field in document analysis due to the low quality and complexity of certain documents [1]. It is, on the one hand, the basis for recognition tasks such as word spotting or OCR. On the other hand, text line detection is used for document structure extraction. Considering printed documents with simple layouts, text line extraction can be solved with simple techniques such as smearing or projection profiles [1]. However, in the field of handwritten documents, text lines may be skewed, touching and contain local distortions. Hence, sophisticated methodologies need to be applied.

Likforman-Sulem et al. [1] published a survey about text line segmentation of historical documents. Well known methodologies for text line extraction are Projection Profiles (PP), Local Projection Profiles [2], [3] and Hough Transform.

More recently, Saabni et al. [4] employ seam carving. They find the minimal costs for a text line by means of dynamic programming. A different approach is proposed by Garz et al. [5] which clusters local features in order to determine text lines in historical manuscripts. They use seam carving so that touching text lines are segmented correctly.

Roy et al. [6] propose a generalized text line extraction for graphical documents. They cluster foreground elements in a hierarchical way such that text lines having multiple orientations or even circular text elements are segmented. An approach using approximated anisotropic Gaussian filter banks targeting camera distorted printed documents is presented by Bukhari et al. [7].

The best performing method at the handwriting segmentation contest (2009 and 2010) is proposed by Shi et al. [8]. They combine steerable directional filters with an Adaptive Local Connectivity Map (ALCM). Lemaitre et al. [9] proposed a top-down approach for text line extraction. They first coarsely detect text lines whose locations are then refined and re-segmented if overlapping connected components are present. Papavassiliou et al. [10] split the document image into vertical zones. An HMM is then used to refine coarsely segmented text and gap regions.

We propose a text line detection methodology which is applicable for handwritten, printed and historical documents. It was initially designed for layout analysis of torn documents with heterogeneous document structures. Subsequent routines use the text lines to detect the page layout or to correctly reassemble document pieces. In order to simplify the interface for these routines, text lines are represented by rectangles whose exact orientation – even if the text line has slight local distortions – is needed for correct alignments. The proposed approach first groups characters to words by means of LPPs. Then, text lines are detected by globally minimizing all word distances.

The proposed approach was evaluated on the most recent Handwriting Segmentation Contests from ICDAR 2009 [11] and ICFHR 2010 [12]. These contests are particularly challenging as they comprise handwriting from different languages, writing systems and writers. The text lines are skewed, contain local distortions and have changing sizes as well as interline spacing. The proposed approach keeps up with state-of-the-art techniques achieving an FM score of 97.06% and 98.59% on the ICFHR 2010 and ICDAR 2009 dataset respectively. Additionally, the methodology was evaluated on the *Saint Gall database*, a historic manuscript. There, background clutter, decorating elements, initials and side notes are present which complicate text line extraction (see Figure 1). On this dataset we improve the current state-of-the-art by gaining an FM score of 99.06%.

The paper is organized as follows. First we present the methodology in Section II. In Section III the performance evaluation on the ICDAR 2009, ICFHR 2010 and Saint Gall database is depicted. Finally, we give a conclusion in Section IV.

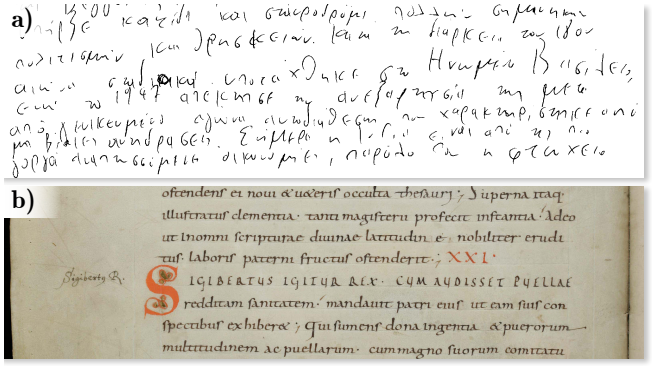

Figure 1. A detail of a sample page from the ICFHR 2010 contest (a) and the Saint Gall database (b).

II. Methodology

The text line detection presented is designed to deal with documents having varying orientations, backgrounds and heterogeneous layouts. First, the documents are aligned using a robust skew detection [13]. Then, the document image is binarized to locate text elements. The binarization method [14] is a scale-space adaption of Su's method [15] which can handle varying font sizes. Subsequently, the connected components are merged to "words" which are then clustered to text lines.

A. Word Estimation

Having aligned and binarized the document images, foreground elements are grouped to words using Local Projection Profiles (LPP). The words detected are then classified into noise, machine printed and handwritten text. The feature extraction and classification is detailed in [16]. Classifying foreground elements has two advantages: First, the labels can be changed from one document scenario to another. Hence, for modern documents we use classes such as machine printed, handwritten, pictures and for historical documents text, comments and initials are distinguished. Second, a binarization that partitions foreground and background elements based on their gradients and gray value is not capable of rejecting unwanted elements such as decorations, graphics or background clutter. That is why an additional foreground class is trained that removes unwanted foreground elements before they spoil the text line detection.

We represent words by means of rectangles so that the text line detection's speed is improved since four values need to be considered rather than the contour or all pixels of a word. Bounding boxes do not account for local variations of a word's orientation. Thus, minimum area rectangles were used in our previous implementation [16]. However, minimum area rectangles are sensitive to non-uniformly distributed ascenders and descenders. Additionally, they cover the whole word rather than the corpus size (x-height), which we are interested in (see Figure 2).

We propose a new abstraction method called *profile boxes* which compensates the drawbacks discussed. Therefore we coarsely detect words by means of oriented bounding boxes. These rectangles are aligned to the document's main orientation which is determined by the skew estimation. Having extracted the binarized word, the upper and lower profiles are computed. Each of these profiles is observed individually and, by means of regression, a line is fitted to the respective profile. Due to the ascenders and descenders, the error distribution of the extracted profiles rather corresponds to heavy-tailed distributions than to normal distributions. Hence, a robust line fitting based on the Welsch distance [17] is performed that compensates for large residual outliers:

$$\rho = min \sum_{i=0}^{n} \frac{C^2}{2} \left(1 - e^{-\left(\frac{r_i}{C}\right)^2}\right) \quad (1)$$

where $C = 2.98$ and r_i are the residuals. Even though this regression method is robust, it may fail if background clutter is present or for short words such as *"to"*. In order to handle such exceptions, the angles of the lines fitted are examined. If their difference is $\leq 5°$, a correct line fitting is assumed and a rectangle whose orientation corresponds to the mean orientation of both lines is constructed. Otherwise, the line with the minimal angle distance to the main orientation is considered for the profile box computation. Solely if both lines fail (their angle difference is $> 5°$), the oriented bounding box is used for representing the word observed.

Figure 2 shows a profile box and the minimum area rectangle respectively. Now that the words are represented by means of profile boxes and classified using Gradient Shape Features (GSF), the text line detection is performed.

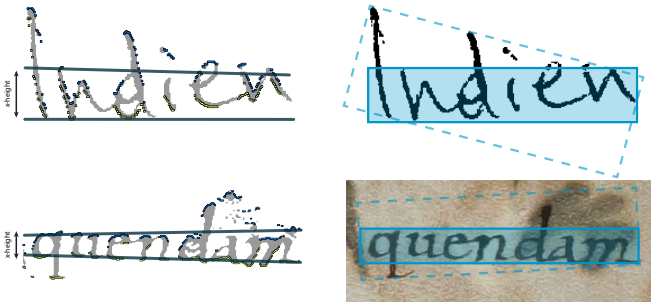

Figure 2. Upper (blue) and lower profiles (yellow dots) used for estimating the profile lines (left). Profile boxes and the corresponding minimum area rectangles (dashed line) which would be larger with a strongly differing orientation from the words' orientations.

B. Text Line Detection

The text line detection presented is a bottom-up approach which utilizes profile boxes of words as basic entity. These words are merged according to their minimal distance. Then, line rectangles are calculated using PCA for a robust line orientation estimation. A recti linearity hypothesis test removes words at either end of a text line that impair the

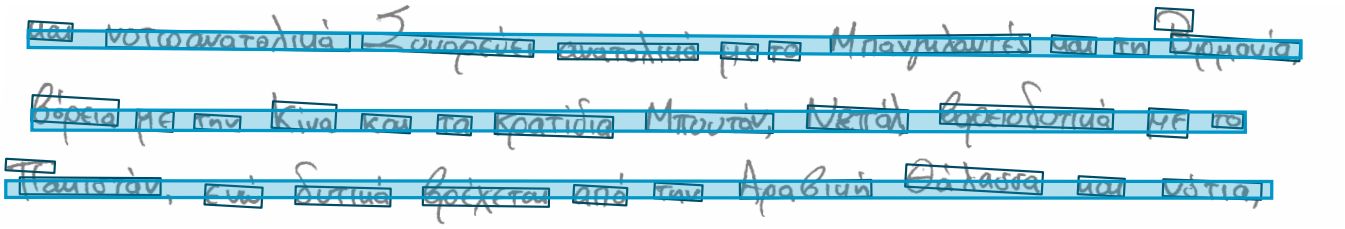

Figure 4. Text line estimation. The last text line is estimated correctly, due to the recti linearity hypothesis test (note the false box above the Π).

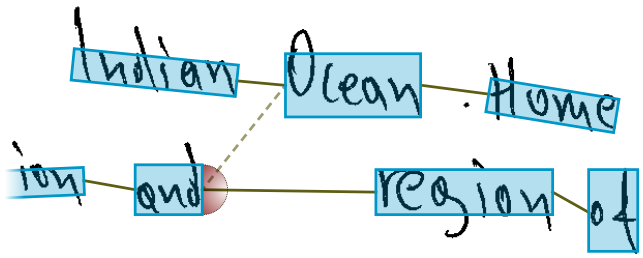

Figure 3. Word merging. The red semi-circle shows the angle weighting proposed. The dashed line illustrates a possible false word merge.

result due to mutually inconsistent locations. Having located the text lines, a back propagation voting is applied which joins local classification results with global knowledge of the current document. Thus, a single word classified as handwriting in the context of printed words may change its label to print depending on its classification accuracies.

One assumption is incorporated into the text line detection: Each word is either the last word of a text line or it is succeeded by exactly one word. This applies to all horizontal and can be easily adapted to vertical writing systems. It implies that maximally two words are merged at the same time and noise in the text region (e.g. quotes or dots) are not merged with any text line. Furthermore, larger irregular spaces between words can be bridged resulting in fewer split errors.

As previously mentioned, the inputs for the text line detection are labeled profile boxes which represent entities like words. First, a distance matrix containing connecting distances $d(p,q)$ of all rectangles is created. Note that "left" and "right" refer to directions normalized by the text's main orientation. Hence, the detection process is carried out in a rotationally invariant manner.

In order to minimize the chance of false vertical merges, caused by low line spacing, the connecting distance of two rectangles is weighted by:

$$d(p,q) = e(p,q) \cdot (1 + C\ g_\sigma(cos(\theta_1 - \theta_2))) \qquad (2)$$

where $e(p,q)$ is the minimal Euclidean distance between the rectangles' left p and right q points (upper, middle and lower). θ_1 is the angle of the left rectangle and θ_2 the angle of the line which connects both rectangles. $g_\sigma(x)$ is a Gaussian distribution with $\sigma = 1/3$ and C is a constant which weights the angle distance depending on the rectangle's class label (based on experiments, we chose $C = 15$ for printed and $C = 4$ for handwritten text). Figure 3 shows a sample image where the Euclidean distance would lead to a false word merging (dashed line), while the distance measure $d(p,q)$ proposed correctly merges the words. Note that in this case *"and"* and *"Ocean"* are not merged anyway because of the global minimization.

Having created the distance matrix, all words with a minimal distance are connected. The decision of connecting neighboring words is based on the global distance minimization. In the end solely words at the beginning or end of a line and noise within the line are left. In order to reject these connections, either a global threshold or an adaptive threshold calculated by means of robust statistics on the distance matrix can be applied.

The orientation of the text line rectangle is estimated using PCA (see Figure 4). First, the upper and lower line of all words merged to one text line are sampled equidistantly so that larger words have greater significance. The PCA is then computed for all sample points. Finally, the line's orientation is determined by:

$$\theta_l = tan^{-1}(e_1, e_2) \qquad (3)$$

where θ_l represents the resulting text line angle and e_1, e_2 are the first and the second Eigenvectors respectively.

III. Results

The text line detection is evaluated on the most recent handwriting segmentation contests from the ICDAR 2009 and ICFHR 2010. Additionally, an evaluation on a dataset with historic manuscripts was performed. These results show, that the proposed approach is applicable for poorly preserved documents. All datasets are written by different writers in English, French, German, Greek and Latin. Hence, the methodology presented is neither depending on the language, the writing systems nor on the writer.

A. Page Segmentation Competition

In order to compare the proposed method with the state-of-the-art in page segmentation, we evaluated it on the last two Page Segmentation Contests of the ICFHR [12] and ICDAR [11]. The datasets consist of 100, respectively

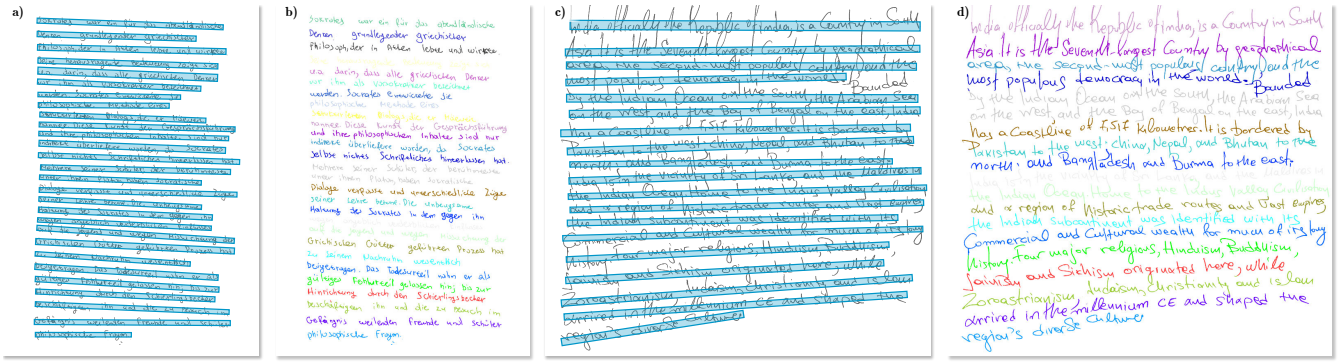

Figure 5. Sample images from the ICDAR 2009 (a, b) and ICFHR 2010 (c, d) datasets. The missing word *"Bounded"* in (c) gets corrected during labeling (d). The FM of (c) is 63.16% by reason of falsely labeled ascenders and descenders especially in the first four text lines (d).

200 handwritten text images written by several authors in English, French, German and Greek. The line spacing is varying, text lines may overlap (e.g. ascenders and descenders) and they are slanted.

The performance metric is based on a MatchScore [12] that computes the maximum overlap of a text region with the ground truth region. If this score is above a given threshold T_α (which is 95% for text line detection), the text line is considered as correct (o2o). Based on this MatchScore, the Detection Rate (DR), the Recognition Accuracy (RA) and the Performance Metric (FM) are computed:

$$DR = \frac{o2o}{N}, \quad RA = \frac{o2o}{M}, \quad FM = \frac{2\ DR\ RA}{DR + RA} \qquad (4)$$

where N is the number of ground truth text lines and M is the number of resulting elements. In other words, the DR can be considered as recall and the RA as precision.

Since our methodology aims at detecting text lines (size, angle, location) rather than perfectly segmenting binarized text, we had to develop a labeling strategy that maps the rectangles to binarized text. In order to achieve this, the word rectangles of each text line are mapped to the binary image with unique line IDs. Then, a region growing guarantees that the ID of the nearest region is mapped to unlabeled elements. By this means, merged ascenders and descenders get different labels. However, the labeling is not optimal for such scenarios (see Figure 5 (d)).

The ICFHR Page Segmentation Contest consists of 100 handwritten images that have $N = 1629$ text lines. The description of the methodologies compared, can be either found in Section I or in [12]. Table I shows, that the proposed method (CVL) can compete with the state-of-the-art, having an FM of 97.06%. The table presents all methods, sorted according to the FM measure, which participated at the ICFHR 2010 Page Segmentation Contest.

The ICDAR 2009 Page Segmentation Contest has 200 images with $N = 4034$ text lines. On this dataset, the text line detection proposed (CVL) performed better with an FM of 98.59% (see Table II).

	M	o2o	DR	RA	FM
CUBS	1626	1589	97.54	97.72	97.63
NifiSoft	1634	1589	97.54	97.25	97.40
CVL	1633	1583	97.18	96.94	97.06
IRISA	1636	1578	96.87	96.45	96.66
ILSP-a	1656	1567	96.19	94.63	95.40
ILSP-b	1655	1559	95.70	94.20	94.95
TEI	1637	1549	95.09	94.62	94.86

Table I
ICFHR 2010 PAGE SEGMENTATION CONTEST [12].

	M	o2o	DR	RA	FM
CUBS	4036	4016	99.55	99.50	99.53
ILSP-LWSeg-09	4043	4000	99.16	98.94	99.05
CVL	4034	3977	98.59	98.59	98.59
PAIS	4031	3973	98.49	98.56	98.52
CMM	4044	3975	98.54	98.29	98.42
CASIA-MSTSeg	4049	3867	95.86	95.51	95.68
PortoUniv	4028	3811	94.47	94.61	94.54
PPSL	4084	3792	94.00	92.85	93.42
LRDE	4423	3901	96.70	88.20	92.25
JadavpurUniv	4075	3541	87.78	86.90	87.34
ETS	4033	3496	86.66	86.68	86.67
AegeanUniv	4054	3130	77.59	77.21	77.40

Table II
ICDAR 2009 PAGE SEGMENTATION CONTEST [11].

B. Saint Gall Database

The second evaluation is carried out on the test set of the *Saint Gall database*. This dataset consists of 30 medieval manuscript pages which are composed of 720 text lines. Pages of this dataset have stains, holes and notes beside the actual text. It can be seen in Figure 6 that the writing differs substantially from modern handwriting. Additionally, initials which span 2-3 text lines complicate the text line detection (see Figure 1). In order to deal with these challenges, the classifier was trained on text, initials and noise (rather than manuscript, printed and noise) using 20 pages from the training set.

For this dataset, the performance is measured by means of the *Pixel-Level Hit Rate (PHR)* and the FM (also called *Line Accuracy Measure*) which were previously used by Garz et al. [5]. The PHR is the number of correctly labeled

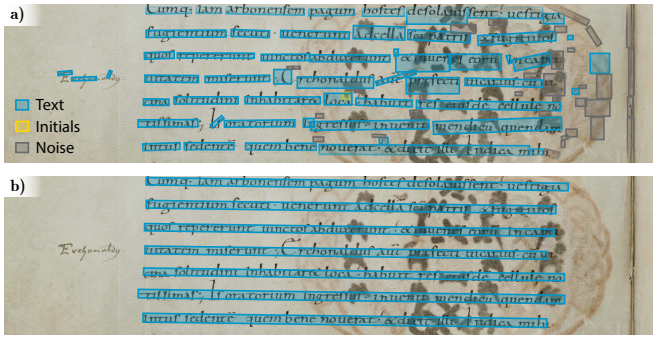

Figure 6. Sample from the Saint Gall database with annotated profile boxes (a) and the resulting text lines (b). Note that the boxes in (a) at the right border are rejected since they are classified as noise.

pixels divided by the number of ground truth pixels. We achieve a PHR of 0.9894 which is a marginal increase of 0.29% compared to previous methods. The proposed method achieves an FM of 0.9903 if the threshold T_α is set to 90% increasing the FM by 1.06% compared to [5].

The evaluations show on the one side that the proposed method is highly adoptable to changing writing styles and page layouts. This results from the bottom-up approach and the flexible word merging which does not incorporate any assumptions of the text line structure. On the other side, in view of the pixel hit rates, one can see that the simple labeling strategy falsely splits connected ascenders and descenders which reduces the performance in such evaluation scenarios (see Figure 5).

IV. Conclusion

We presented a text line detection methodology that allows for recognizing text lines independent to the language, writing system or writer. The bottom-up approach incorporates solely one assumption about the text lines present, namely that a word has exactly one successor. Using profile boxes instead of minimum area rectangles or bounding boxes improves the word localization while still representing words in a compact manner. In addition, this abstraction layer significantly speeds up the processing time as the computation is carried out on hundreds of words rather than millions of pixels (it takes $2.241sec$ on an average page of the ICFHR 2010 competition with a 3 GHz Dual Core processor). With a simple labeling strategy, the text line rectangles can be mapped to connected components in a post-processing step which allows for a text line segmentation on pixel basis.

The methodology presented is especially suitable for text line detection scenarios where the document layout varies substantially. Even though it was designed for modern documents with printed and handwritten text, we applied it successfully to handwritten manuscripts and historic documents. The evaluation showed, that it can compete with state-of-the-art methods without incorporating high-level machine learning such as HMMs or dataset dependent knowledge.

Acknowledgment

The authors would like to thank Dirk Phler, Jan Schneider and the Fraunhofer-Institute for Production Systems and Design Technology (IPK), Berlin for supporting the work.

References

[1] L. Likforman-Sulem, A. Zahour, and B. Taconet, "Text Line Segmentation of Historical Documents: A Survey," *IJDAR*, vol. 9, no. 2-4, pp. 123–138, 2007.

[2] A. Alaei, U. Pal, and P. Nagabhushan, "A New Scheme for Unconstrained Handwritten Text-Line Segmentation," *Pattern Recognition*, vol. 44, no. 4, pp. 917–928, 2011.

[3] Y. Gao, X. Ding, and C. Liu, "A Multi-scale Text Line Segmentation Method in Freestyle Handwritten Documents," in *ICDAR*, 2011, pp. 643–647.

[4] R. Saabni and J. El-Sana, "Language-Independent Text Lines Extraction Using Seam Carving," in *ICDAR*, 2011, pp. 563–568.

[5] A. Garz, A. Fischer, R. Sablatnig, and H. Bunke, "Binarization-Free Text Line Segmentation for Historical Documents Based on Interest Point Clustering," in *Document Analysis Systems*. IEEE, 2012, pp. 95–99.

[6] P. P. Roy, U. Pal, and J. Lladós, "Text line extraction in graphical documents using background and foreground information," *IJDAR*, vol. 15, no. 3, pp. 227–241, 2012.

[7] S. S. Bukhari, F. Shafait, and T. M. Breuel, "Text-Line Extraction Using a Convolution of Isotropic Gaussian Filter with a Set of Line Filters," in *ICDAR*, 2011, pp. 579–583.

[8] Z. Shi, S. Setlur, and V. Govindaraju, "A Steerable Directional Local Profile Technique for Extraction of Handwritten Arabic Text Lines," in *ICDAR*, 2009, pp. 176–180.

[9] A. Lemaitre, J. Camillerapp, and B. Coüasnon, "A perceptive method for handwritten text segmentation," in *DRR*, ser. SPIE Proceedings, vol. 7874. SPIE, 2011, pp. 1–10.

[10] V. Papavassiliou, T. Stafylakis, V. Katsouros, and G. Carayannis, "Handwritten document image segmentation into text lines and words," *Pattern Recognition*, vol. 43, no. 1, pp. 369–377, 2010.

[11] B. Gatos, N. Stamatopoulos, and G. Louloudis, "ICDAR 2009 Handwriting Segmentation Contest," in *ICDAR*, 2009, pp. 1393–1397.

[12] ——, "ICFHR 2010 Handwriting Segmentation Contest," in *ICFHR*, 2010, pp. 737–742.

[13] M. Diem, F. Kleber, and R. Sablatnig, "Skew Estimation of Sparsely Inscribed Document Fragments," in *Document Analysis Systems*, 2012, pp. 292–296.

[14] F. Kleber, M. Diem, and R. Sablatnig, "Scale Space Binarization Using Edge Information Weighted by a Foreground Estimation," in *ICDAR*, 2011, pp. 854–858.

[15] B. Su, S. Lu, and C. L. Tan, "Binarization of historical document images using the local maximum and minimum," in *Document Analysis Systems*. ACM, 2010, pp. 159–166.

[16] M. Diem, F. Kleber, and R. Sablatnig, "Text Classification and Document Layout Analysis of Torn Documents," in *ICDAR*, 2011, pp. 1181–1184.

[17] R. E. Welsch and E. Kuh, "Linear Regression Diagnostics," Massachusetts Institute of Technology, Tech. Rep. 923-77, April 1977.