High Performance Layout Analysis of Medieval European Document Images

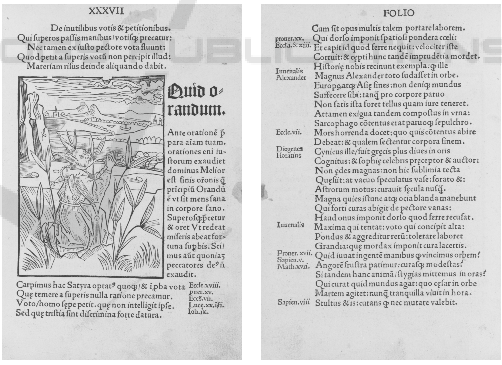
Syed Saqib Bukhari[1], Ashutosh Gupta[1,2], Anil Kumar Tiwari[2] and Andreas Dengel[1,3]
[1]*German Research Center for Artificial Intelligence, Kaiserslautern, Germany*
[2]*IITJ-Indian Institute of Technology Jodhpur, India*
[3]*Technical University Kaiserslautern, Germany*

Keywords: Document Analysis, Historical Document Analysis, Layout Analysis, Document Image Segmentation.

Abstract: Layout analysis, mainly including binarization and page segmentation, is one of the most important performance determining steps of an OCR system for complex medieval document images, which contain noise, distortions and irregular layouts. In this paper, we present high performance page segmentation techniques for medieval European document images which include a novel main-body and side-notes segregation and an improved version of OCRopus (OCRopus,) based text line extraction. In order to complete the high performance layout analysis pipeline, we have also presented the application of the percentile based binarization (Afzal et al., 2014) and the multiresolution morphology based text and non-text segmentation (Bukhari et al., 2011) methods over historical document images. presented layout analysis techniques are applied to a collection of the 15th century Latin document images, which achieved more than 90% accuracy for each of the segmentation techniques.

1 INTRODUCTION

This paper addresses the problem of layout analysis of historical European document images. Most languages of Europe belong to the Indo-European language family. This family is divided into a number of branches, including Romance, Germanic, Baltic, Slavic, Albanian, Celtic, Armenian and Hellenic (Greek). The Uralic languages, which include Hungarian, Finnish, and Estonian, also have a significant presence in Europe. Example of Latin European document images are shown in Figure1.



XXXVII

De inutilibus votis & petitionibus.
Qui ſuperos paſſis manibus/votiſq; precatur:
Nec tamen ex iuſto pectore vota fluunt:
Quod petit a ſuperis votũ non percipit illud:
Materiam riſus deinde aliquando dabit.

Quid orandum.

Ante oratione p̃
para aiam tuam.
orationes eni iu-
ſtorum exaudiet
dominus Melior
eſt finis oronis q̃
pricipiũ Orandũ
ē vt ſit mens ſana
in corpore ſano.
Superoſq; p̃cetur
& oret Vt redeat
miſeris abeat for-
tuna ſupbis. Sci/
mus aũt quoniã
peccatores de9 ñ
exaudit.

Carpimus hac Satyra optat9 quoq;/& ipba vota Eccle.xviii.
Quę temere a ſuperis nulla ratione precamur. puer.xv.
Voto/homo ſępe petit. quę non intelligit ipſe. Eccs.vii.
Sed quę triſtia ſint diſcrimina forte datura. Lucę.xx.iſi.
Ioh.ix.

(a)

FOLIO

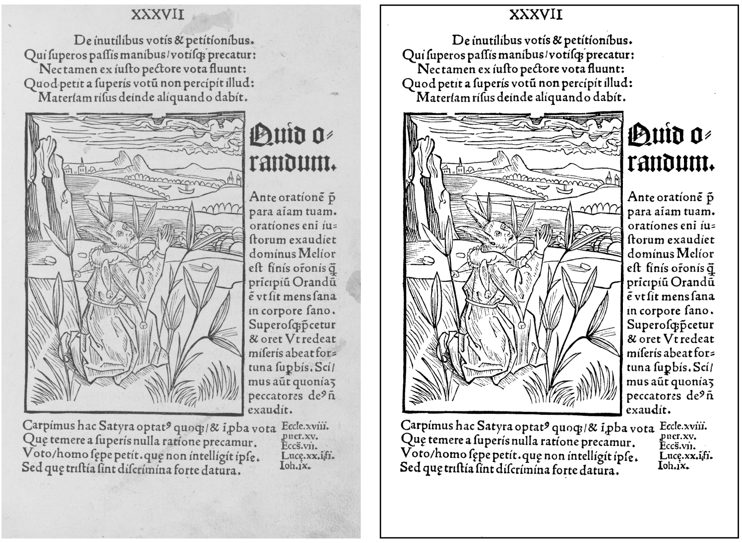
Cum ſit opus multis talem portare laborem.
prouer.xx. Qui dorſo imponit ſpatioſi pondera cœli:
Eccl.i.& xiii. Et capiti id quod ferre nequit: velociter iſte
Corruit: & cępti hunc tandē imprudētia mordet.
Hiſtorię nobis recinunt exempla: ꝙ ille
Iuuenalis Alexander Magnus Alexander toto ſudaſſet in orbe.
Europę atq; Aſię fines: non deniq; mundus
Suffecere ſibi: tanq̃ pro corpore paruo
Non ſatis iſta foret tellus quam iure teneret.
Attamen exigua tandem compoſtus in vrna:
Sarcophago cōtentus erat paruoq; ſepulchro.
Eccle.vii. Mors horrenda docet: quo quis cōtentus abire
Debeat: & qualem ſectentur corpora finem.
Diogenes Horatius Cynicus ille/fuit gręcis plus diues in oris
Cognitus: & ſophię celebris pręceptor & auctor:
Non ędes magnas: non hic ſublimia tecta
Quęſiit: at vacuo ſpeculatus vaſe: forato &:
Aſtrorum motus: curauit ſęcula nuſq̃.
Magna quies iſtunc atq; ocia blanda manebunt
Qui forti curas abigit de pectore vanas:
Haud onus imponit dorſo quod ferre recuſat.
Iuuenalis Maxima qui tentat: voto qui concipit alta:
Pondus & aggreditur rerũ: tolerare laboret
Grandia: quę mordax imponit cura lacertis.
Prouer.xvii. Quid iuuat ingentē manibus ꝙ vincimus orbem?
Sapien.v. Math.xvii. Angorē fruſtra patimur: curaſq; modeſtas?
Si tandem hanc animã/ſtygias mittemus in oras?
Qui curat quid mundus agat: quo cęſar in orbe
Martem agitet: nunq̃ tranquilla viuit in hora.
Sapien.viii Stultus & is: curans ꝙ nec mutare valebit.

(b)

Figure 1: 15^{th} century Medieval European Documents from the Kallimachos Project (Kallimachos,); (a), on the left, contains both Text and Non-Text regions; Document (b), on the right, contains only Text regions.

Layout analysis, text and non-text segmentation, main-body and side-notes segregation, and text-line extraction, is a major performance limiting step in large scale document digitization projects. Over the last two decades, several layout analysis algorithms have been proposed in the literature (Cattoni et al., 1998), (Nagy, 2000) that work for different layouts, scripts and are quite robust to the presence of noise in documents. Here, we briefly discuss some state-of-the-art document image layout analysis approaches in connection to European documents. Text and non-text segmentation is an important layout analysis step, which may directly affect the performance of further layout processing tasks such as text-line extraction, and/or character recognition. The performance of classification based on text and non-text segmentation approaches (Bukhari et al., 2010) heavily depends on training samples, and they can not be directly applied to different scripts. On the other hand, smearing (Wong et al., 1982) and multiresolution morphology (Bloomberg, 1991), (Bukhari et al., 2011) based approaches work on an assumption that non-text el-

Bukhari, S., Gupta, A., Tiwari, A. and Dengel, A.
High Performance Layout Analysis of Medieval European Document Images.
DOI: 10.5220/0006574603240331
In *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2018)*, pages 324-331
ISBN: 978-989-758-276-9



XXXVII

De inutilibus votis & petitionibus.
Qui ſuperos paſſis manibus/votiſq; precatur:
Nec tamen ex iuſto pectore vota fluunt:
Quod petit a ſuperis votũ non percipit illud:
Materiam riſus deinde aliquando dabit.

Quid o-
randum.

Ante orationẽ p̃
para aĩam tuam.
orationes eni iu=
ſtorum exaudiet
dominus Melior
eſt finis orõnis q̃
pricipiũ Orandũ
ẽ vt ſit mens ſana
in corpore ſano.
Superoſq; p̃cetur
& oret Vt redeat
miſeris abeat for=
tuna ſup̱bis. Sci/
mus aũt quoniaʒ
peccatores de⁹ ñ
exaudit.

Carpimus hac Satyra optat⁹ quoq;/& ip̱ba vota Eccle.xviii.
Quę temere a ſuperis nulla ratione precamur. p̃uer.xv.
Voto/homo ſępe petit.quę non intelligit ipſe. Eccš.vii.
Sed quę triſtia ſint diſcrimina forte datura. Lucę.xx.iſi.
Ioh.ix.

XXXVII

De inutilibus votis & petitionibus.
Qui ſuperos paſſis manibus/votiſq; precatur:
Nec tamen ex iuſto pectore vota fluunt:
Quod petit a ſuperis votũ non percipit illud:
Materiam riſus deinde aliquando dabit.

Quid o-
randum.

Ante orationẽ p̃
para aĩam tuam.
orationes eni iu=
ſtorum exaudiet
dominus Melior
eſt finis orõnis q̃
pricipiũ Orandũ
ẽ vt ſit mens ſana
in corpore ſano.
Superoſq; p̃cetur
& oret Vt redeat
miſeris abeat for=
tuna ſup̱bis. Sci/
mus aũt quoniaʒ
peccatores de⁹ ñ
exaudit.

Carpimus hac Satyra optat⁹ quoq;/& ip̱ba vota Eccle.xviii.
Quę temere a ſuperis nulla ratione precamur. p̃uer.xv.
Voto/homo ſępe petit.quę non intelligit ipſe. Eccš.vii.
Sed quę triſtia ſint diſcrimina forte datura. Lucę.xx.iſi.
Ioh.ix.

(a) Scanned Document (b) Binarized Document

Figure 2: The Percentile based Binarization Methodology(Afzal et al., 2014). Input scanned document (a) is binarized using percentile filtering to give binary output document (b).

ements are bigger than text elements, however these approaches are script independent and can be directly used for European script document images.

Text-line extraction is the backbone of a layout analysis system. Kumar et al. (Kumar et al., 2007) have evaluated the performance of six algorithms for page segmentation on Nastaliq script: the x-y cut (Nagy et al., 1992), the smearing (Wong et al., 1982), whitespace analysis (Baird, 1994), the constrained text-line finding (Baird, 2002), Docstrum (OGorman, 1993), and the Voronoi-diagram based approach (Kise et al., 1998). These algorithms work very well in segmenting documents in Latin script as shown in (Shafait et al., 2008). However, none of these algorithms were able to achieve an accuracy of more than 70% on their test data which had simple book layouts. More sophisticated approaches for text-line extraction have been presented in the domain of segmenting handwritten European document so far. However, the key problem addressed in these approaches is to handle local non-linearity of text-lines.

In this paper, we present a high performance layout analysis system for a wide variety of Historical European document images that belong to a diverse collection of layout structures such as books, magazines, and newspapers. Our layout analysis system is a suitable combination of robust and well-established text and non-text segmentation, main-body and side-notes segregation, and text-line extraction techniques. First, it performs text and non-text segmentation using multiresolution morphology based method (Bukhari et al., 2011). Then, it segregates main-body and side-notes based on vertical white space calculation and filtering for a variety of single and multi-column layouts. Finally, it determines the text-lines that are extracted based on y-derivative of Gaussian kernel. In this way, our layout analysis system extends OCRopus (OCRopus,) based layout analysis system (text-line extraction) by incorporating text and non-text segmentation, a novel main-body and side-notes segregation and an improvised text-line extraction method. To evaluate the performance of the presented layout analysis system for real-world documents, a dataset of European scanned documents is prepared. This paper focuses on an extensive experimental evaluation of the presented layout analysis system and its comparison with state-of-the-art techniques. The rest of this paper is organized as follows. Our layout analysis system for historical European document images is described in Section II. Performance evaluation and experimental results are discussed in Section III, followed by a conclusion in Section IV.

2 HIGH PERFORMANCE LAYOUT ANALYSIS OF HISTORICAL EUROPEAN DOCUMENT IMAGES

The high performance layout analysis of historical European document images in this paper comprises of the following main steps; binarization and page segmentation which includes text and non-text segmentation, main body and side note segregation and text-line extraction. For this purpose, more specifically, we have proposed a novel main-body and side-notes segregation technique, and we have improved OCRopus (OCRopus,) based text-line extraction technique. Together with that we have applied the percentile based binarization method (Afzal et al., 2014) of OCRopus and Bukhari et al. (Bukhari et al., 2011) based text and non-text segmentation techniques on historical documents. For the completeness of this paper, together with explaining our novel main-body and side-notes segregation technique and improved version of OCRopus (OCRopus,) based text-line extraction method, we have also briefly described the percentile based binarization method and the multiresolution morphology based text and non-text segmentation (Bukhari et al., 2011) based techniques.

A brief description of these steps is provided here:

2.1 The Percentile based Binarization Method (Afzal et al., 2014)

In general, an image can be thresholded by determining a global threshold for the entire page (known as global binarization method) or by using the statistics obtained from a local window centered around the pixel which is being thresholded. The percentile based binarization method (Afzal et al., 2014) by OCRopus takes into consideration the background statistics based on percentile filters.In this method, text and non-text regions are treated equally for determining the threshold used for local binarization. It works well both on focused and defocused images. This method also works well on monocular images with defocused parts. The binarization method starts with estimating the background at each location in the image,i.e., a whole new image is created having only the background of the image based on percentile.The threshold in this method is adapted in accordance with the background properties of the image. The original image has a domain of all gray level values, i.e., [0,255] and the background image estimated for each value based on percentile filters at every location has a domain of only two levels,i.e.,0,255. The thresholding is done in a way that if the pixel value in original image is less than 't' times the pixel value in background estimated image, then the corresponding pixel value in the output image is labeled one, where 't' is the parameter used to determine that whether a pixel is foreground or background, depending on the similarity of the pixel, and the background, which has been estimated using percentile filter; otherwise it is labeled zero.

2.2 The Multiresolution Morphology based Text and Non-text Segmentation (Bukhari et al., 2011)

Bloomberg (Bloomberg, 1991) presented a multiresolution morphology based text and non-text segmentation method. It is a simple and script independent text and non-text segmentation method. It performs well for halftone mask segmentation, for which it was designed, but most of the time fails to accurately segment drawing type non-text elements such as line art, maps etc. Bukhari et al. (Bukhari et al., 2011) presented an improved multiressolution morphology based text and non-text segmentation algorithm, that can handle halftones as well as drawing type non-text elements. A sample document image and its text and non-text segmentation results for the original and the improved version of multiresolution morphology based methods are shown in Figure 3.

2.3 The Improved Text-Line Extraction Method

The text-line extraction technique proposed here is a modified version of the OCRopus' text-line extraction method, which is called "ocropus-gpageseg". The OCRopus technique of text-line extraction is explained briefly here;

It first estimates the "scale" of the text by finding connected components of individual letters in the binary image and calculating the median of their dimensions. It removes components which are too big or too small (according to scale) which are unlikely to be letters.

In the baseline ocropus-gpageseg method, column separators in binary image are found using convolution and thresholding. At first vertical white spaces on binary image are found and then the rest region is filled in order to form smooth text region using filtering. Then using Guassian and uniform filtering, the column edges (gradients) are found in the binary image by setting a certain threshold in accordance with the scale of the image. Then the smoothened text region and the column edges are combined to get column separators. In the next step, out of the total column separators, only selected number of column separators with dimensions greater than minimum value are selected.

In order to find text lines, at first, box-map (bounding box) is found by setting two thresholds. If the area of the slice list lies in between the threshold areas, then that slice is labeled one, otherwise it is labeled zero- it helps in removing noise. Then a clean

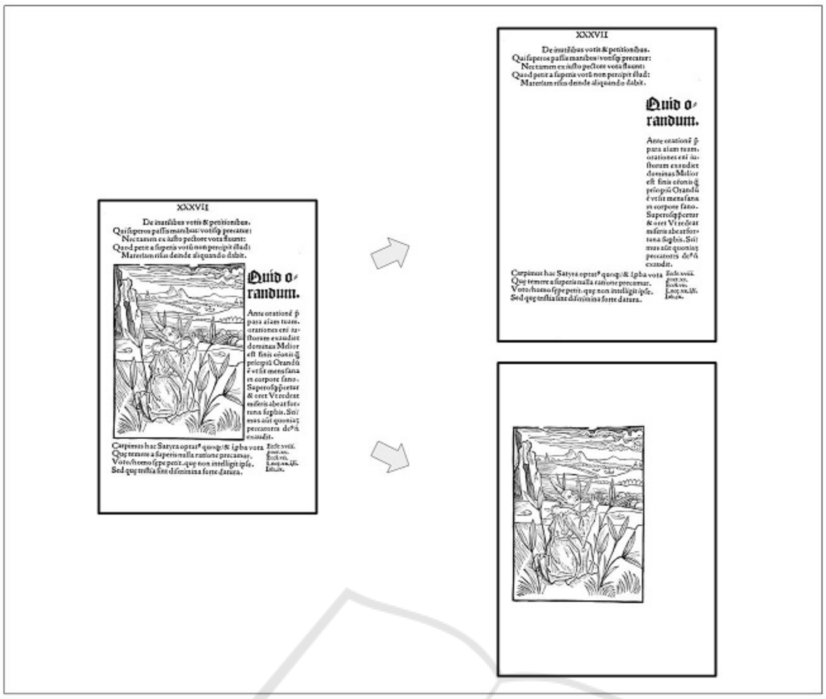


Figure 3: Text & Non-Text Segmentation Methodology by Bukhari et al. (Bukhari et al., 2011). The input image, on the left, is segmented into two images- one containing only Text regions and the other containing only Non-Text regions.

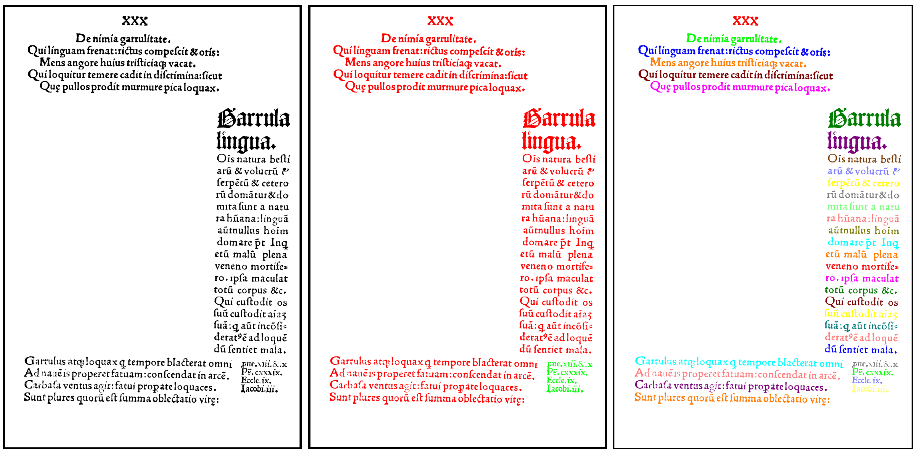
image is obtained by multiplying the two image arrays of box-map and the given binary image, keeping only the desired text. On this cleaned image, the y-derivative of a Gaussian kernel is applied to detect the top and bottom edges of the remaining features. It then blurs this horizontally to blend the tops of letters on the same line together. The same is done with the bottoms of the letters. The areas between top and bottom edges are blurred and treated as text line regions and termed as line seeds.

Then column separators and line seeds are combined and used to segment the binary images. Basically, column separators restrict line seeds i.e., separate two lines horizontally.

In our presented modified version of ocropus-gpageseg method, column separators in binary image are found more accurately using convolution and thresholding with some optimal parametrical changes and post-processing steps like removal of two column separators that are too close to each other in the same horizontal line and extension of column separators to the first and last rows of the image with a condition that no character is crossed in between on the extended path. At first vertical white spaces on binary image are found and then the rest region is labeled in order to form smooth text region using filtering. Then using Gaussian and uniform filtering, the column edges (gradients) are found in the binary image by setting a certain threshold in accordance with the scale of the image. The smoothened text region and the column edges are combined to get column separators. In the next step, out of the total column separators, only selected number of column separators with dimension greater than min value are selected. The finally obtained column separators are then combined with the initially obtained text region (through white space method) in order to find more precise text only regions in the binary image. All the gaps/holes within the text regions are filled up and thus final text only regions are obtained. In the Improved OCRopus Text-Line Method, we focused more on extracting the precise text only regions which form a sentence and separate them from other text regions like side-notes which are too close to the main-body text regions and then extract text lines using y-derivative of Gaussian kernel and filtering. The result of improved OCRopus text-line extraction method is shown in 4.

2.4 The Novel Main Body and Side Notes Segregation Technique

In this segregation technique, the main objective is the classification of text only regions after segment-



(a) (b) (c)

Figure 4: Image(a): Binarized European Document; Image(b): Main-Body and Side-Notes Segregated Document; Image(c): Improved OCRopus based Text-Line Segmented Document.

ing the binary image of a European document into text and non-text. After removing the non-text regions and major noise content from the binary image, the image is smoothened in order to label the text regions and form a blob over them by finding vertical white spaces and applying Gaussian and uniform filtering.The blobs are formed over the text regions in such a way that the text-lines which are not part of a sentence but appear too close to each other are also separated. Among all the blobs formed over text regions, the ones below a certain adaptive threshold width are classified as side notes and the rest are labeled as main body text regions. Too small blobs below a certain adaptive threshold with respect to median of heights of every character connected component is considered as noise and hence removed. The result of main-body and side-notes segregation method is shown in 4.

3 PERFORMANCE EVALUATION

The 15th century novel “Narrenschiff” is part of the German government funded project Kallimachos (Kallimachos,). For the performance evaluation of the proposed layout analysis techniques, we have selected a subset of 50 images from one of the Latin novels in the Kallimachos project. Sample document images are shown in Figure 1. These images contain both text and non-text regions, as well as main body and side notes within text regions. For this dataset, text and non-text segmentation, main body and side note segregation, and text-line extraction ground-truths containing both text and non-text regions are prepared in color coded pixel form as shown in Figures 6,7 and 8. The images have variety of both single and multi-column layouts and hence they can be used to evaluate the performance of a layout analysis algorithms for European document images. Below, the performance evaluation of the presented layout analysis techniques is done in three parts. The first part evaluates the performance of text and non-text segmentation, the second part analyzes the errors made in main body and side note segregation, and the third part evaluates the overall accuracy of text-line extraction technique.

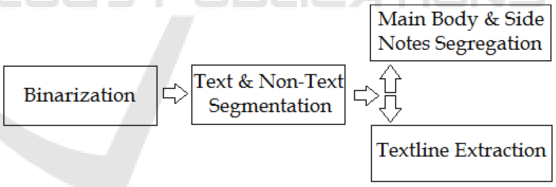

Binarization
Text & Non-Text Segmentation
Main Body & Side Notes Segregation
Textline Extraction

Figure 5: Complete Methodology.

3.1 Text and Non-text Segmentation

As stated above, our dataset contains 50 historical document images. We test the performance of our approach using images with different writing styles and layout structures which were not used for training.

Pixel-level ground truth has been generated by manually assigning text in the documents of the testing set with one of the two classes, main-body or side-notes text. Several methods to measure the segmentation accuracy have been reported in literature. We evaluate the segmentation accuracy by adopting the F-measure metric which combines precision and recall values into a single scalar representative. It guarantees that both values are high (conservative), in con-

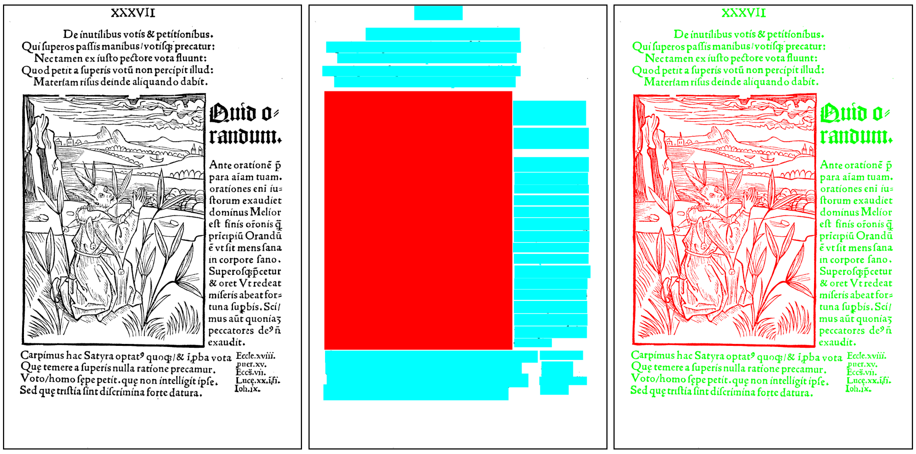


Figure 6: Text & Non-Text Ground Truth Generation.

trary to the average (tolerant) which does not hold this property. For example, when precision and recall both equals one, the average and F-measure will both be one, but, if the precision is one and the recall is zero, the average would be 0.5 and the F-measure would be zero. Therefore,this measure has been adopted as it reliably measures the segmentation accuracy. Precision and recall are estimated according to Eq. 1 and Eq. 2, resp.

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

where *True-Positive(TP)*, *False-Positive(FP)*and *False-Negative(FN)* with respect to side-notes, are defined as following:

- TP:side-notes text classified as side-notes text.
- FP:side-notes text classified as main-body text.
- FN:main-body text classified as side-notes text.

Likewise, these metrics can also be defined with respect to main-body text. Once we have the precision and recall counts, F-measure is calculated according to Eq. 3.

$$F-Measure = \frac{(1+\beta^2)*Precision*Recall}{\beta^2*Recall+Precision} \quad (3)$$

Assigning $\beta = 1$ induces equal emphasis of precision and recall on F-measure estimation. The F-Measure accuracies are shown in Table 1.

F-measure for both main-body and side-notes text with different postprocessing window sizes is shown in Table 1. Note that the optimal window size is 100.

Table 1: Performance of Text and Non-Text Extraction method by calculating F-Measure for Text and Non-Text regions.

	Text and Non-Text Segmentation
Main-Body F-Measure(%)	99.433%
Side-Notes F-Measure(%)	99.6683%

3.2 Main Body and Side Note Segregation

The performance evaluation matrices for main body and side note segregation accuracy are based on f-measure calculation as described in previous section. The F-Measure accuracies are shown in Table 2.

Table 2: Performance of Main-Body and Side-Notes segregation method by calculating F-Measure for main body and side note text regions.

	Main-Body and Side-Notes Segregation
Main-Body F-Measure(%)	99.7646%
Side-Notes F-Measure(%)	80.4962%

3.3 Text-line Extraction

The ground-truth images for evaluation of Text-Line Extraction Technique performance are created manually by pixel coloring

The performance evaluation metrics for text-line detection accuracy are defined in (Shafait et al., 2008), where a text-line is said to be correctly detected if it does not fall into any of the following categories of errors: over-segmentation, under-segmentation, missed text-lines, and false-alarms. Let,N_g:ground-truth text-lines;N_s:segmented text-lines;N_{o2o}:one-2-one correctly detected text-lines. The one-to-one text-line detection accuracy is represented by Eq. 4.

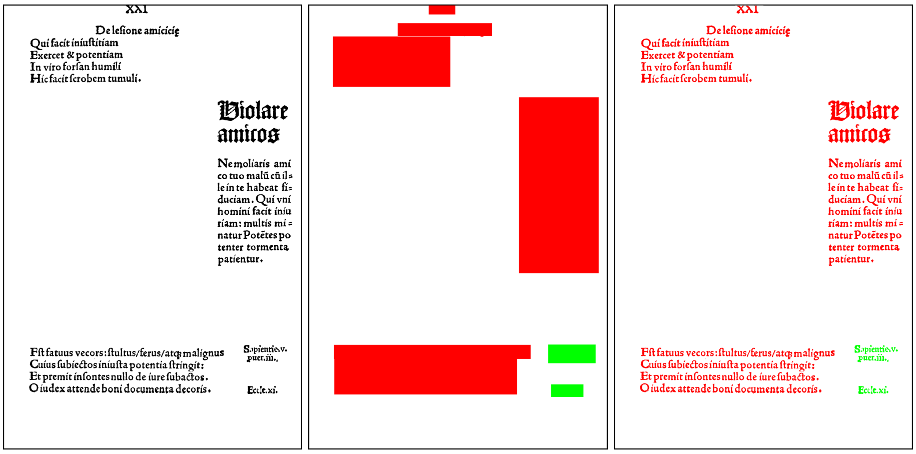


Figure 7: Ground-Truth Generation for Main & Side Body Segregation.

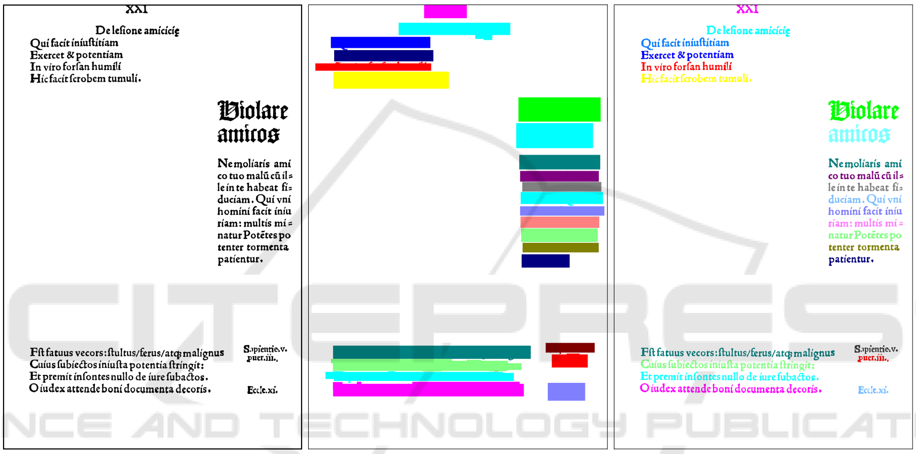


Figure 8: Ground-Truth Generation for Text-Line Extraction.

$$P_{o2o}\% = \frac{N_{o2o}}{N_g} \qquad (4)$$

For modified text-line extraction methodology on European dataset, we achieved a performance gain from 72.18% to 94.53% after text and non-text segmentation as shown in Table 3.

Table 3: Performance of Improved Text-Line Extraction method based on performance evaluation metrics for text-line detection accuracy defined in (Shafait et al., 2008).

Technique	Accuracy(%)
OCRopus-gpageseg	72.177338%
Improved OCRopus-gpageseg	94.530014%

4 CONCLUSION

In this paper, we have presented a high performance layout analysis system for historical European document images, which are composed of a variety of single and multi-column layouts. The presented layout analysis system is composed of a suitable combination of well-established and robust text and non-text segmentation, novel main-body and side-notes segregation, and text-line extraction methods. We have evaluated the presented layout analysis system on the dataset of 50 document images from a 15th century Latin script historical novel from the Kallimachos project (Kallimachos,), which are composed of a different layout structures as shown in Figure 1 containg both text and non-text regions. For text and non-text segmentation, multiresolution morphology based method (Bukhari et al., 2011) is used. We have achieved above 99%text and non-text segmentation accuracy on the dataset. For main-body and side-notes segregation, the methodology is explained in (Section II-D). For this mthod, we achieved 99% main-body segregation accuracy and above 80% side-notes accuracy for the dataset. For text-line extraction, improved version of OCRopus based text-line extraction method is used, which is described in (Section II-C). For Improved OCRopus based text-line extraction method, we have achieved above 94% text-line extraction accuracy for the dataset, which is better than the performance of above 72% of OCRopus-gpageseg method on the dataset. Altogether, the presented layout analysis system showed good performance for text and non-text segmentation, main-body and side-notes segregation, and text-line extraction on a variety of European document images, and it can be used for large scale European documents digitization processes.

ACKNOWLEDGEMENTS

This work was partially funded by the BMBF (German Federal Ministry of Education and Research), project Kallimachos (01UG1415C).

REFERENCES

Afzal, M. Z., Krämer, M., Bukhari, S. S., Yousefi, M. R., Shafait, F., and Breuel, T. M. (2014). Robust binarization of stereo and monocular document images using percentile filter. In *Revised Selected Papers of the International Workshop on Camera-Based Document Analysis and Recognition - Volume 8357*, pages 139–149, New York, NY, USA.

Baird, H. S. (1994). Background structure in document images. In *in Document Image Analysis, H. Bunke, P. Wang, and H. S. Baird, Eds. World Scientific, Singapore*.

Baird, H. S. (2002). Two geometric algorithms for layout analysis. In *Proc. Workshop on Document Analysis Systems*.

Bloomberg, D. S. (1991). Multiresolution morphological approach to document image analysis. In *Proc. International Conference on Document Analysis and Recognition, Franc*.

Bukhari, S. S., Shafait, F., and Breuel, T. (2010). Document image segmentation using discriminative learning over connected components. In *Proc. Workshop on Document Analysis Systems, Boston, US*.

Bukhari, S. S., Shafait, F., and Breuel, T. (2011). Improved document image segmentation algorithm using multiresolution morphology. In *SPIE, Document Recognition and Retrieval XVIII*.

Cattoni, R., Coianiz, T., Messelodi, S., and Modena, C. M. (1998). Geometric layout analysis techniques for document image understanding: a review. In *IRST, Trento, Italy, Tech. Rep*.

Kallimachos. www.kallimachos.de.

Kise, K., Sato, A., and Iwata, M. (1998). Segmentation of page images using the area voronoi diagram. In *Computer Vision and Image Understanding*.

Kumar, K. S., Kumar, S., and Jawahar, C. (2007). On segmentation of documents in complex scripts. In *Proc. International Conference on Document Analysis and Recognition*.

Nagy, G. (2000). Twenty years of document image analysis in pami. In *IEEE Trans. on Pattern Analysis and Machine Intelligencen*.

Nagy, G., Seth, S., and Viswanathan, M. (1992). A prototype document image analysis system for technical journals. In *Computer, vol. 7, no. 2*.

OCRopus. https://github.com/tmbdev/ocropy.

OGorman, L. (1993). The document spectrum for page layout analysis. In *IEEE Trans. on Pattern Analysis and Machine Intelligencen*.

Shafait, F., Keysers, D., and Breuel, T. M. (2008). Performance evaluation and benchmarking of six page segmentation algorithms. In *IEEE Trans. on Pattern Analysis and Machine Intelligencen*.

Wong, K. Y., Casey, R. G., , and Wahl, F. M. (1982). Document analysis system. In *IBM Journal of Research and Development*.