Document Image Segmentation using Discriminative Learning over Connected Components

Syed Saqib Bukhari
Technical University of Kaiserslautern, Germany
bukhari@informatik.uni-kl.de

Mayce Ibrahim Ali Al Azawi
Technical University of Kaiserslautern, Germany
ali@iupr.com

Faisal Shafait
German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany
faisal.shafait@dfki.de

Thomas M. Breuel
Technical University of Kaiserslautern, Kaiserslautern, Germany
tmb@informatik.uni-kl.de

ABSTRACT

Segmentation of a document image into text and non-text regions is an important preprocessing step for a variety of document image analysis tasks, like improving OCR, document compression etc. Most of the state-of-the-art document image segmentation approaches perform segmentation using pixel-based or zone(block)-based classification. Pixel-based classification approaches are time consuming, whereas block-based methods heavily depend on the accuracy of block segmentation step. In contrast to the state-of-the-art document image segmentation approaches, our segmentation approach introduces connected component based classification, thereby not requiring a block segmentation beforehand. Here we train a self-tunable multi-layer perceptron (MLP) classifier for distinguishing between text and non-text connected components using shape and context information as a feature vector. Experimental results prove the effectiveness of our proposed algorithm. We have evaluated our method on subset of UW-III, ICDAR 2009 page segmentation competition test images and circuit diagrams datasets and compared its results with the state-of-the-art leptonica's [1] page segmentation algorithm.

1. INTRODUCTION

Document image segmentation is the problem of classifying the contents of a document image into a set of text and non-text classes. Non-text class consists of following categories: halftone, drawing, maths, logos, tables, etc. Document image segmentation is one of the most important preprocessing steps before feeding the specific contents to an optical character recognition (OCR) system otherwise OCR engine produces lot of garbage characters originated from non-text components, as shown in Figure 1.

Document image segmentation approaches in the literature can generally be classified into two groups: (i) block or zone based classification and (ii) pixels based classification. Block based segmentation approaches apply page segmentation [11] on the document image and then classify the obtained blocks into a set of determined classes [5]. On the other hand pixel based approaches attempt to classify individual pixels [8, 7] according to predefined classes.

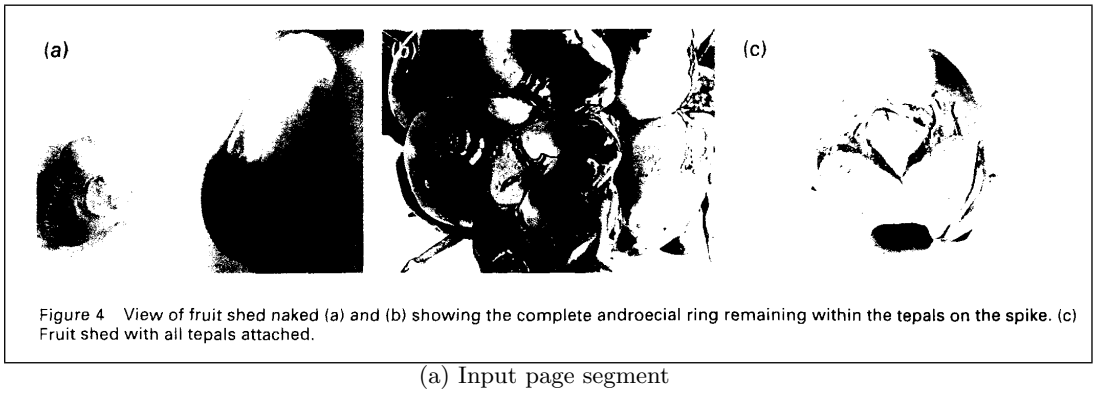
Several block classification algorithms have been proposed over the years. For a more detailed overview of related work in the field of document block classification please refer to Okun [9] and Wang [12]. Okun et al. [9] proposed an approach for document block classification based on connected components and run-length statistics. Wang et al. [12] presented the block classification system, each block with a 25 dimensional feature vector and use an optimized decision tree classifier to classify each block into one of different target classes. The most recent and detailed block classification approach is introduced by Keysers et. al [5] which showed that a document block classification system can be constructed using run-length histogram feature vector alone. That work includes several classes of blocks (math, logo, text, table, drawing, halftone, ruling and speckles). In general, the approaches that classify blocks depend heavily on the result of page segmentation into blocks. The blocks may be segmented in a wrong way leading to miss-classification.

Moll et al. [8, 7] classify individual pixels instead of regions, to avoid the constriction of the limited classes of region shapes. The approach is applied on handwritten, machine printed and photographed document images. Pixel based classification approaches are slow with respect to execution time. The approach by Won [13] focuses on a combination of a block based algorithm and a pixel based algorithm to segment a document image into text and image area.

Together with block based and pixel based image segmentation approaches, there is another state-of-the-art text and

[1] http://code.google.com/p/leptonica/

DAS '10, June 9-11, 2010, Boston, MA, USA



(a) Input page segment

tv _ _. - l 4
la) ` )âo‹ , ` (C)
Â» ` fr /.1 Â·r , ` WM. âo›
7 if Â¢âo Â¥, ` _` { Ã©>Ã©_ j l; I X I I
K;~Â· - _V . Â· ` _ ~ M" JF Q} { fÂ·',; Ai ' V _ I \' g,
âo _ it ~ _y, i * t
f jgj . __ it ` l âo› f
1< .Â· âo› Â·. V Â·- âo‹Â¢ âo‹ âo‹ âo› F {U
Â·~ . 7 ~ Ã© . i/`I / _âo› * âo‹ K Â»
Figure 4 View of fruit shed naked la) and lb) showing the complete androecial ring remaining within the tepals on the spike. (c) Fruit shed with all tepals attached.

(b) OCR result

Figure 1: The OCR result of an in-correctly segmented zone containing both images and text. The OCR system generates many garbage symbols from the non-text parts of the input page segment.

halftone segmentation approach reported by Bloomberg et al. [2] based on multi-resolution morphological operations. This approach comprises three steps: 1) at first step, seed image is generated by sub-sampling input image such that the resulting seed image mainly contains halftone pixels. 2) Then mask image is produced by using morphological operations such that together with all image pixels there is a sufficient connectivity of halftone seed pixels with other pixels covering halftone regions. 3) In last step binary filling operation is used to transform seed image with the help of mask image into final halftone mask image. The open-source version of this algorithm is presented in leptonica library developed by Dan Bloomberg. This method produces promising results for halftone objects but is unable to recognize thin halftone and drawing like objects as non-text objects.

In this paper our aim is to perform text and non-text classification based on connected components, instead of pixels or blocks. For this purpose, we use simple and easy to compute feature vectors. For training we use multi-layer perception (MLP) classifier which has already been used in different document image pre-processing tasks [6], like binarization [4], deskewing [10]. Classifier tuning is considered as one of the hard problem with respect to the optimization of parameters. In order to get rid of this problem we use self-tunable MLP classifier [3]. Our method is independent of block segmentation and equally applicable to different categories of non-text objects if they were included in the training data. One can analyze the ease of implementation and accuracy of our approach in algorithm description and experimental evaluation sections respectively.

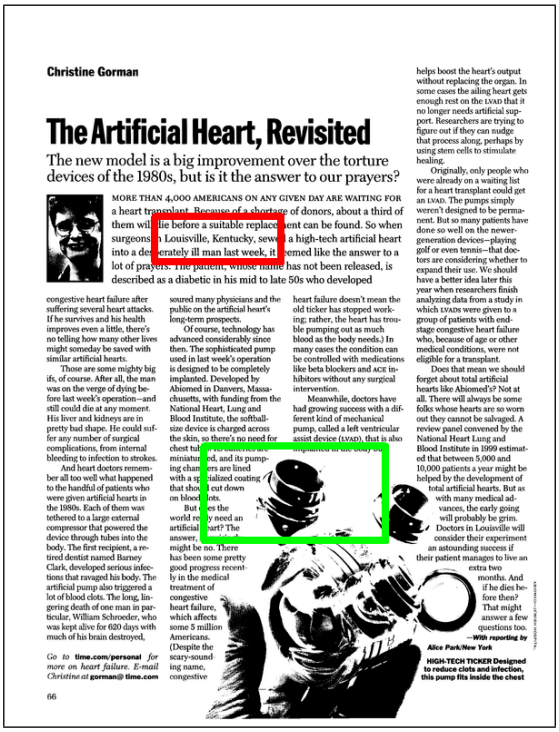
The rest of this paper is organized as follows. In Section 2 we describe our document image segmentation method. Section 3 deals with the experimental results. Section 4 describes conclusion.

2. DOCUMENT IMAGE SEGMENTATION ALGORITHM

Here we describe our document image segmentation algorithm which segments document image into text and non-text regions. Our main target is to classify each connected component as either text or non-text component. In Section 2.1 we describe feature extraction process. In Section 2.2 we discuss about the training of extracted features using self-tunable multi-layer perceptron (MLP) classifier.

2.1 Feature Extraction

Instead of extracting complex features from a connected component, the raw shape of a connected component itself is an important distinguishable feature for classifying structured text and random or irregular non-text components, as shown in Figure 2. Together with the shape of connected component, the surrounding area of a connected component can also play an important role for text and non-text classification, similarly because of the structured text and non-structured non-text surrounding areas. Figure 2 shows neighborhood surrounding areas for text and non-text regions. We refer connected component with its neighborhood surrounding as context. Based on the above mentioned hypothesis, our feature vector of connected component is composed of shape and context information. Detail description of the feature vector is presented below.



Christine Gorman

The Artificial Heart, Revisited

The new model is a big improvement over the torture devices of the 1980s, but is it the answer to our prayers?

MORE THAN 4,000 AMERICANS ON ANY GIVEN DAY ARE WAITING FOR a heart transplant. Because of a shortage of donors, about a third of them will die before a suitable replacement can be found. So when surgeons in Louisville, Kentucky, sewed a high-tech artificial heart into a desperately ill man last week, it seemed like the answer to a lot of prayers. The patient, whose name has not been released, is described as a diabetic in his mid to late 50s who developed congestive heart failure after suffering several heart attacks. If he survives and his health improves even a little, there's no telling how many other lives might someday be saved with similar artificial hearts.

Those are some mighty big ifs, of course. After all, the man was on the verge of dying before last week's operation—and still could die at any moment. His liver and kidneys are in pretty bad shape. He could suffer any number of surgical complications, from internal bleeding to infection to strokes.

And heart doctors remember all too well what happened to the handful of patients who were given artificial hearts in the 1980s. Each of them was tethered to a large external compressor that powered the device through tubes into the body. The first recipient, a retired dentist named Barney Clark, developed serious infections that ravaged his body. The artificial pump also triggered a lot of blood clots. The long, lingering death of one man in particular, William Schroeder, who was kept alive for 620 days with much of his brain destroyed, soured many physicians and the public on the artificial heart's long-term prospects.

Of course, technology has advanced considerably since then. The sophisticated pump used in last week's operation is designed to be completely implanted. Developed by Abiomed in Danvers, Massachusetts, with funding from the National Heart, Lung and Blood Institute, the softball-size device is charged across the skin, so there's no need for chest tubes. Its components are miniaturized, and its pumping chambers are lined with a special coating that should cut down on blood clots.

But does the world really need an artificial heart? The answer, it might be no. There has been some pretty good progress recently in the medical treatment of congestive heart failure, which affects some 5 million Americans. (Despite the scary-sound ing name, congestive heart failure doesn't mean the old ticker has stopped working; rather, the heart has trouble pumping out as much blood as the body needs.) In many cases the condition can be controlled with medications like beta blockers and ACE inhibitors without any surgical intervention.

Meanwhile, doctors have had growing success with a different kind of mechanical pump, called a left ventricular assist device (LVAD), that is also implanted in the body and helps boost the heart's output without replacing the organ. In some cases the ailing heart gets enough rest on the LVAD that it no longer needs artificial support. Researchers are trying to figure out if they can nudge that process along, perhaps by using stem cells to stimulate healing.

Originally, only people who were already on a waiting list for a heart transplant could get an LVAD. The pumps simply weren't designed to be permanent. But so many patients have done so well on the newer-generation devices—playing golf or even tennis—that doctors are considering whether to expand their use. We should have a better idea later this year when researchers finish analyzing data from a study in which LVADs were given to a group of patients with end-stage congestive heart failure who, because of age or other medical conditions, were not eligible for a transplant.

Does that mean we should forget about total artificial hearts like Abiomed's? Not at all. There will always be some folks whose hearts are so worn out they cannot be salvaged. A review panel convened by the National Heart Lung and Blood Institute in 1999 estimated that between 5,000 and 10,000 patients a year might be helped by the development of total artificial hearts. But as with many medical advances, the early going will probably be grim. Doctors in Louisville will consider their experiment an astounding success if their patient manages to live an extra two months. And if he does, before then? That might answer a few questions too.

—With reporting by Alice Park/New York

Go to **time.com/personal** *for more on heart failure. E-mail Christine at* **gorman@time.com**

HIGH-TECH TICKER Designed to reduce clots and infection, this pump fits inside the chest

66

Figure 2: Sample image from ICDAR 2009 page segmentation competition. This image shows the structured shapes of text components and random shapes of non-text components.

- **shape of connected component**: In document images, most of the text components are smaller than non-text components. Therefore size information can play an important role in the text and non-text components classification. But only size information is not enough for classifying the big text and the small non-text components. Therefore, together with size information we need some other features as well. As already mentioned, the shapes of non-text connected components are irregular, random and vary a lot form one image to another and on other hand the shapes of text components are uniformly structured in document images. The structured and random shapes of text and non-text components respectively can be learned by the MLP classifier. For generating feature vector, each connected component is rescaled to a 40×40 pixel window size. This rescaling performs only downscaling, such that a connected component is downscaled if either length or height of component is greater than 40 pixels otherwise it is fit into the center of a 40×40 window. The advantage of doing this type of rescaling is to distinguish the shape of small components from large components. This type of rescaling can produce different feature vectors for a same components, for example small and big font 'a'. Our target is not to classify each characters but to classify the text and non-text components. Therefore, in our case this rescaling works better than normal rescaling because of incorporating implicit size information of text and non-text components. Rescaled text and non-text connected components are shown in Figure 3(a) and Figure 3(b) respectively. Together with raw rescaled connected component, our shape based feature vector is also composed of four other size based features, mentioned below. So all together the size of our shape-based feature vector is 1604.

  1. normalized length (length of a component divided by the length of an input image).
  2. normalized height (height of a component divided by the height of an input image).
  3. aspect ratio of a component (length divided by height).
  4. number of foreground pixels in a rescaled area divided by total rescaled area.

- **surrounding context of connected component**: Usually the text components are aligned horizontally in the document images which results in structured surrounding area for a text component as compared to the non-structured surrounding area for non-text components. Therefore, the surrounding context of a connected component can play an important role in classifying the text and the non-text components. Each connected component with its surrounding context area is rescaled to a 40×40 window size for generating context-based feature vector. Here the surrounding context area is not fixed for all of the connected components for calculating feature vectors, but it is a function of component's length(l) and height(h). Such that, for each connected component the area of dimensions $5 \times l$ by $2 \times h$ is chosen empirically by keeping a connected component at center for rescaling. The rescaled text and non-text context components are shown in Figure 3(c) and Figure 3(d) respectively. The size of context-based feature vector is 1600.

In this way, the size of a complete feature vector is 3204 which consist of raw rescaled shape (dimension 1600), raw rescaled context (dimension 1600) and four size based features.

2.2 Classification

In general, classifier tuning is a hard problem with respect to the optimization of their sensitive parameters, for example learning rate of MLP classifier, 'C' and gamma of SVM classifier, confidence of decision tree classifier, maximum depth and number of attributes of random forest classifier, 'k' of K nearest neighbor classifier etc. We use MLP classifier for text and non-text classification. Performance of MLP classifier is sensitive to the chosen parameters values. The optimal parameters values depend upon the dataset. The parameters optimization problem can be solved by using grid search for classifier training. But grid search is a slow process. Therefore in order to overcome this problem we use AutoMLP [3], a self-tuning classifier that can automatically adjust learning parameters. In AutoMLP classifier we train a population for MLP classifiers in parallel. For these MLP classifiers, learning parameters are selected from parameter space which has been sampled according to some probability distribution function. All of these MLPs are trained for few

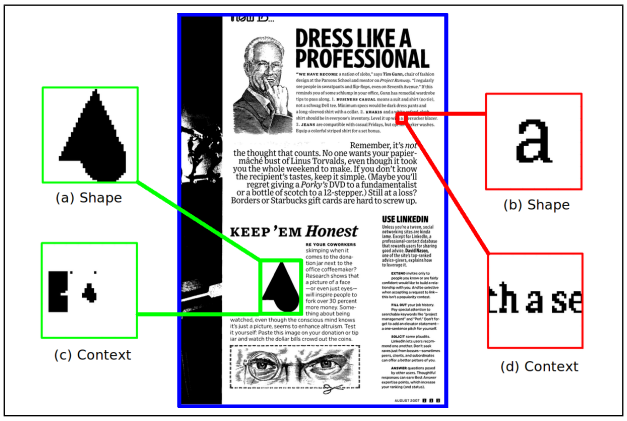


Figure 3: Text and non-text connected components shape and context features, (a) and (b) show rescaled (no upscaling, either downscale or fit into the center to preserve size) connected component's shape features. (c) and (d) show rescaled connected component's context features.

epochs and then half of these classifiers are selected for next generation based on the better performance. The AutoMLP performs internal validation on a portion of training data.

Feature vectors for training AutoMLP classifier have been extracted from the UW-III dataset. The UW-III dataset contains zone-level ground truth for text, halftone, ruling, drawing and logo. From this zone-level ground-truth information, the text and the non-text (halftone, drawing and logo) regions are extracted form document images. Non-text regions were small in number, which have been increased up to four times by rotating each non-text region in four different orientations. Around 0.7 million text samples and 0.1 million non-text samples are used for training AutoMLP classifier.

For testing and evaluation purpose, the feature vector for each connected component of a test document image is extracted in the same way as described in Section 2.1. Then a class label is assigned to each connected component based on classification probabilities of text and non-text.

In order to improve the segmentation results, a nearest neighbor analysis by using class probabilities is performed for refining the class label of each connected component. For this purpose, a region of 70×70 (empirically chosen) is selected from document image by keeping targeted connected component at center. The probabilities of connected components within the selected regions are already computed during classification. Already assigned class labels of the connected components are updated using the average text and non-text provabilities of connected components within selected region. Some of segmented results are shown in Figure 4.

3. EXPERIMENTAL RESULTS

We have evaluated our document image segmentation approach using UW-III, ICDAR-2009 page segmentation competition test dataset [1] and our private circuit diagrams datasets. The main reason for using different datasets is to check the accuracy of our approach on different types of images which have not been used in training as well as to have a variety of text and non-text components. For example, majority of the document images in UW-III dataset have Manhattan-layout but ICDAR 2009 dataset also contains documents with non-Manhattan layout. All non-text components, except halftone, have been removed from UW-III and ICDAR-2009 test datasets. In contrast to this, the circuit diagrams dataset mainly composed of text and drawing components having no other types of non-text components. Total 95 documents have been selected from UW-III dataset. ICDAR 2009 dataset contains 8 test images. Our circuit diagrams dataset composed of 10 images.

For each dataset, pixel-level ground truth has been generated using zone-level ground truth information. Each pixel in ground-truth images contains either text or non-text label. Different types of metrics have been used for the performance evaluation of document image segmentation method which are defined below:

1. **non-text classified as non-text:** percentage of intersection of non-text pixels in both segmented and ground truth image with respect to the total number of non-text pixels in ground truth image.

2. **non-text classified as text:** percentage of intersection of text pixels in segmented image and non-text pixels in ground truth image with respect to the total number of non-text pixels in ground truth image.

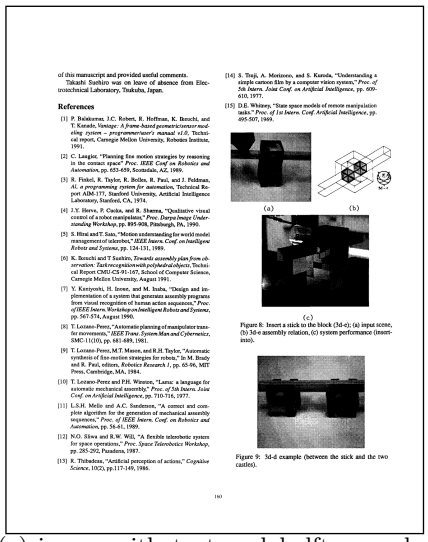
3. **text classified as text:** percentage of intersection of text pixels in both segmented and ground truth image with respect to the total number of text pixels in ground truth image.

4. **text classified as non-text:** percentage of intersection of non-text pixels in segmented image and text pixels in ground truth image with respect to the total number of text pixels in ground truth image.

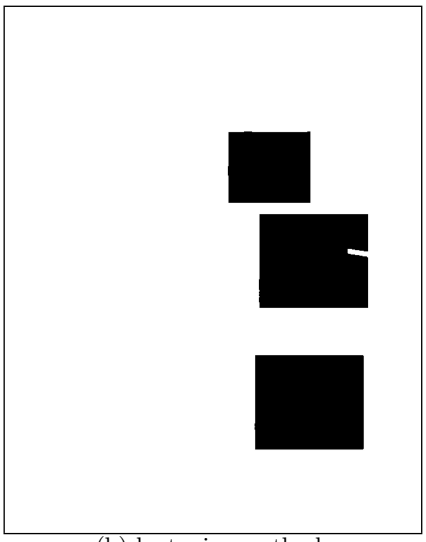
5. **segmentation accuracy:** average percentage of text classified as text accuracy and non-text classified as non-text accuracy.

Based on the matrices defined above, we have compared our approach with leptonica's page-segmentation algorithm. Leptonica algorithm is exclusively designed for segmenting text and halftone components. Performance comparison results of our and leptonica methods on UW-III and ICDAR 2009 test datasets which contains only text and halftone components are shown in Table 1. The boxplot of text and halftone accuracy of our and leptonica methods on combined UW-III and ICDAR 2009 test datasets is shown in Figure 5. Our algorithm has also been evaluated on circuit diagrams dataset in order to show its potential as compared to other text and halftone based segmentation approaches like leptonica, results are shown in Table 2.

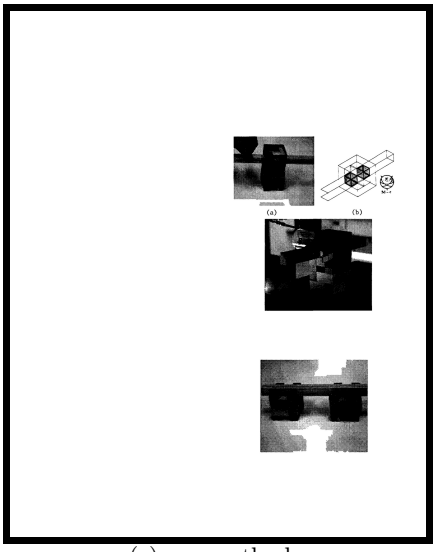
4. DISCUSSION



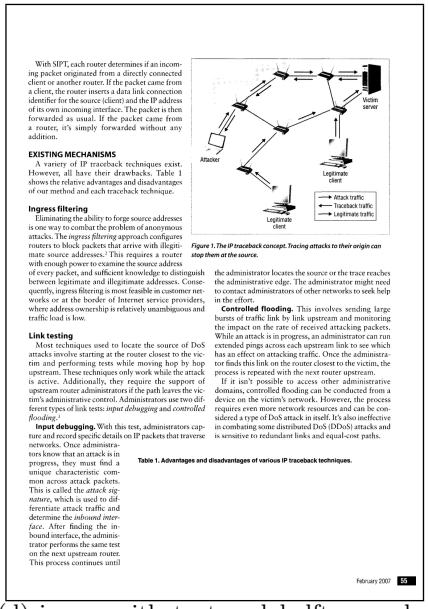
(a) image with text and halftone only (UW-III)



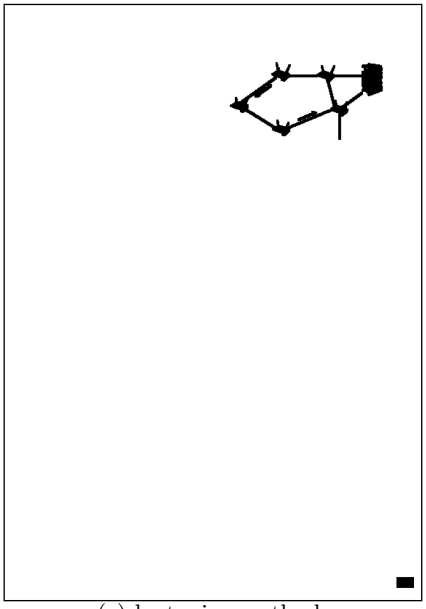
(b) leptonica method



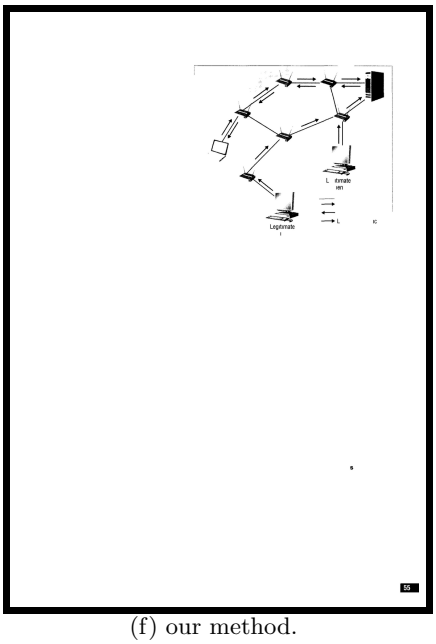
(c) our method.



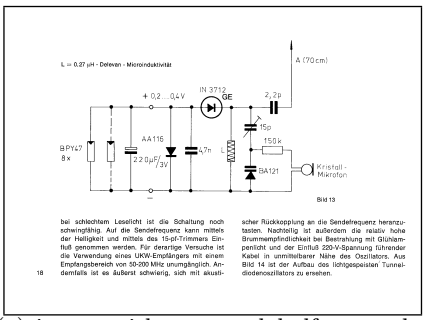
(d) image with text and halftone only (ICDAR 2009)



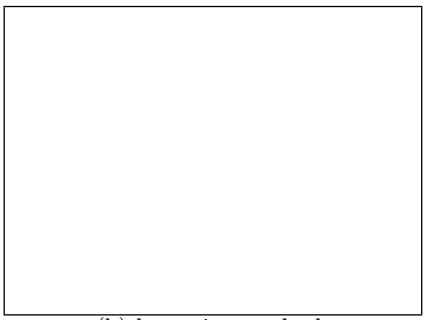
(e) leptonica method



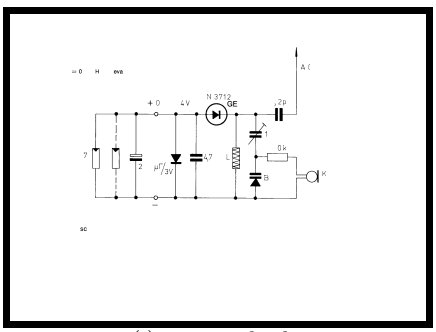
(f) our method.



(g) image with text and halftone only (Circuit Diagram)



(h) leptonica method



(i) our method.

Figure 4: Document image segmentation results of our and leptonica methods in non-text mask format.

Table 1: Performance evaluation of our and leptonica page segmentation algorithms on UW-III dataset (95 document images), ICDAR 2009 page segmentation competition test dataset (8 document images) and combined UW-III and ICDAR 2009 datasets (103 document images).

	UW-III		ICDAR-2009		*Combined*	
	our approach	leptonica	our approach	leptonica	*our approach*	*leptonica*
non-text classified as non-text	98.91%	95.36%	96.70%	84.91%	*98.79%*	*94.77%*
non-text classified as text	1.09%	4.64%	3.30%	15.09%	*1.21%*	*5.23%*
text classified as text	95.93%	99.79%	93.31%	99.87%	*95.72%*	*99.79%*
text classified as non-text	4.07%	0.21%	6.69%	0.13%	*4.28%*	*0.21%*
segmentation accuracy	97.42%	97.57%	95.01%	92.39%	*97.25%*	*97.28%*

Table 2: Performance evaluation results of our and leptonica page segmentation algorithms on circuit diagrams dataset (10 document images). Note: Leptonica method is designed for text and halftone segmentation. Here we used it for evaluating it on circuit diagrams dataset to show that, unlike our method, usually text and halftone based segmentation methods can not be directly applied on other types of non-text components segmentation.

	our approach	leptonica
non-text classified as non-text	89.79%	0%
non-text classified as text	10.21%	100%
text classified as text	89.29%	100%
text classified as non-text	10.72%	0%
segmentation accuracy	89.54%	50%

We have described and experimentally evaluated a new method for document image segmentation into text and non-text regions based on discriminative learning over connected components. We have used the self-tuning MLP classifier (AutoMLP) [3] which automatically optimized learning parameters. Our method is independent of preprocessing step of zone segmentation which is usually the case in zone based classification approaches. We have evaluated our algorithm on UW-III, ICDAR 2009 page segmentation competition test dataset and circuit diagrams datasets and compared its results with state-of-the-art leptonica's page segmentation method [2]. In general, both the text and non-text components are equally important in document image analysis operations. For example, OCR exclusively requires text components and the document image compression or symbol recognition approaches exclusively require non-text components. The performance evaluation results of our and leptonica methods are presented in Table 1, Table 2 and Figure 5. It is obvious from the results that leptonica method has better text classification accuracy than non-text classification. Leptonica method miss classifies the small non-text components as the text components, as shown in Figure 4(b) and Figure 4(e). On the other hand, our method gives equal importance to both the text and non-text components during the classification. Unlike leptonica method, our method can also classify between the small non-text and text components, as shown in Figure 4(c) and Figure 4(f). Leptonica method is designed for the text and halftone segmentation and is not specifically designed for the drawing objects segmentation. Therefore, it is unable to recognize drawing images in circuit diagram dataset, as shown in Table 2 and Figure 4(h). Together with halftone components segmentation, our method also has a potential of segmenting drawing components (for example circuit diagrams), as shown in Table 2 and Figure 4(i). The segmentation results of our method can be improved by increasing training samples and/or by using some post-processing operations.

5. ACKNOWLEDGMENTS

This work was partially funded by the BMBF (German Federal Ministry of Education and Research), project PaREn (01 IW 07001).

6. REFERENCES

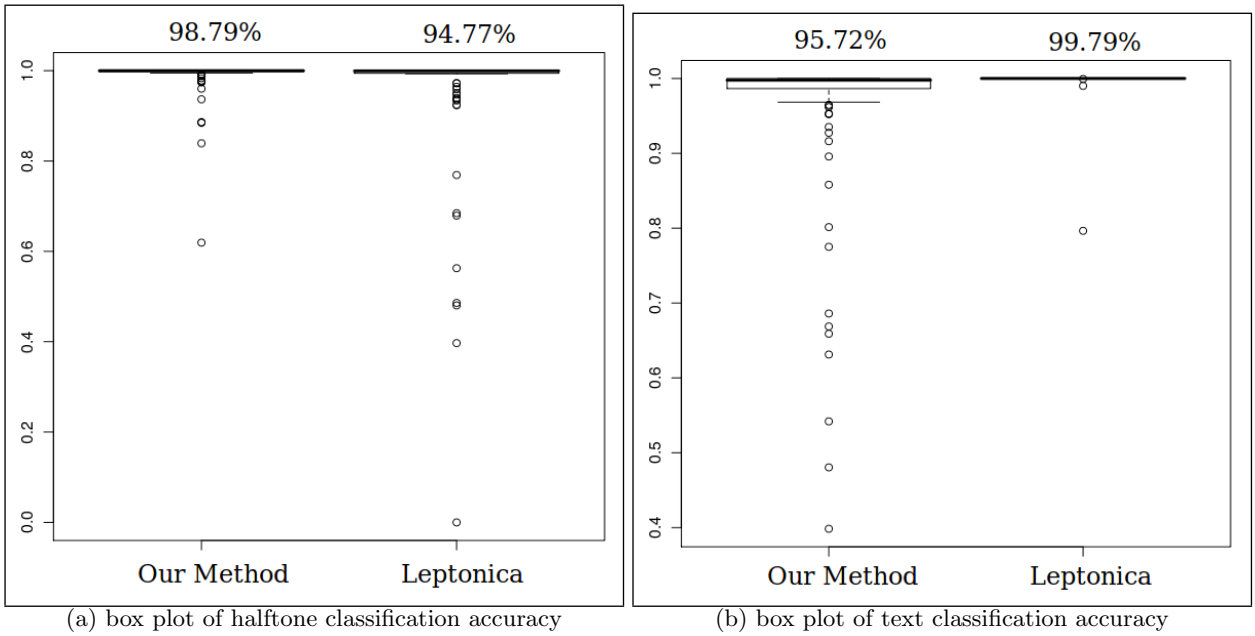
[1] A. Antonacopoulos, D. Bridson, C. Papadopoulos, and S. Pletschacher. ICDAR 2009 page segmentation competition. In *Proc. Int. Conf. Documnet Analysis and Recognition (ICDAR2009)*, pages 1370–1374, Barcelona, Spain, 2009.

[2] D. S. Bloomberg and F. R. Chen. Extraction of text-related features for condensing image documents. In *SPIE Conf. 2660, Document Recgnition III*, pages 72–88, San Jose, CA, 1996.

[3] T. M. Breuel and F. Shafait. Automlp: Simple, effective, fully automated learning rate and size adjustment. In *The Learning Workshop, Snowbird, Utah*, 2010.

[4] Z. Chi and K. W. Wong. A two-stage binarization approach for document images. In *Proc. Int. Symp. Intelligent Multimedia, Video and Speech Processing (ISIMP'01)*, pages 275–278, 2001.

[5] D. Keysers, F. Shafait, and T. M. Breuel. Document



(a) box plot of halftone classification accuracy

(b) box plot of text classification accuracy

Figure 5: Box Plots of our and leptonica page segmentation algorithms on 103 images of UW-III and ICDAR-2009 page segmentation competition test datasets. Average classification accuracies are shown on the top of boxplots.

image zone classification- a simple high-performance approach. In *Proc. 2nd Int. Conf. Computer Vision Theory and Applications*, pages 44–51, Barcelona, Spain, Mar. 2007.

[6] S. Marinai, M. Gori, and G. Soda. Artificial neural networks for document analysis and recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 27(1), Jan. 2005.

[7] M. A. Moll and H. S. Baird. Segmentation-based retrieval of document images from diverse collections. In *Document Recognition and Retrieval XV, Proc. of the SPIE*, volume 6815, pages 68150L–68150L, 2008.

[8] M. A. Moll, H. S. Baird, and C. An. Truthing for pixel-accurate segmentation. In *Document Analysis Systems, the Eighth IAPR Int. Workshop*, pages 379–385, Sep. 2008.

[9] O. Okun, D. Doermann, and M. Pietikainen. Page segmentation and zone calssification: the state of art. In *Technical Report LAM-TR-036, CAR-TR-927, CS-TR-4079*, University of Maryland, College Park, Nov. 1999.

[10] N. Rondel and G. Breuel. Coorperation of multilayer perceptrons for the estimation of skew angle in text document images. Proc. Int. Conf. Documnet Analysis and Recognition (ICDAR'95), pages 1141–1144, 1995.

[11] F. Shafait, D. Keysers, and T. M. Breuel. Performance evaluation and benchmarking of six page segmentation algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(6):941–954, Jun 2008.

[12] Y. Wang, I. Phillips, and R. Haralick. Document zone content classification and its performance evaluation. In *Pattern Recognition*, volume 39, pages 57–73, 2006.

[13] C. S. Won. Image extraction in digital documents. In *Journal of Electronic Imaging*, volume 17, page 033016, 2008.