Stamp Detection in Color Document Images

Barbora Micenková[1,2]
[1]*Department of Computer Science*
Aarhus University (AU)
Aarhus, Denmark
Barbora.Micenkova@dfki.de

Joost van Beusekom[2]
[2]*Multimedia Analysis and Data Mining Group*
German Research Center for Artificial Intelligence (DFKI)
Kaiserslautern, Germany
Joost.van_Beusekom@dfki.de

***Abstract*—An automatic system for stamp segmentation and further verification is needed especially for environments like insurance companies where a huge volume of documents is processed daily. However, detection of a general stamp is not a trivial task as it can have different shapes and colors and, moreover, it can be imprinted with a variable quality and rotation. Previous methods were restricted to detection of stamps of particular shapes or colors. The method presented in the paper includes segmentation of the image by color clustering and subsequent classification of candidate solutions by geometrical and color-related features. The approach allows for differentiation of stamps from other color objects in the document such as logos or texts. For the purpose of evaluation, a data set of 400 document images has been collected, annotated and made public. With the proposed method, recall of 83% and precision of 84% have been achieved.**

***Keywords*-stamp detection; image segmentation; computational forensics; color clustering**

I. Introduction

Despite an enormous utilization of computer technology in various areas of our lives, paper documents still play an important role. Contracts, wills, certificates, invoices and all documents issued by formal authorities are printed on a solid paper and a signature or a stamp guarantee the authenticity of the content. Official stamped documents are often accepted without questioning, omitting the fact that there is an advanced computer technology available to public which can be easily misused for fraud.

The process of modern forgery involves either photocopying or scanning the original document, digital image modification and printing. A huge volume of documents is processed daily in offices such as insurance companies or banks and the degree of automation is still increasing. For example, the printed invoices incoming to an insurance company are immediately digitized and there is no time and resources to manually check if the stamp is authentic [1], which makes the forgery rather easy. For that purpose, an automatic system is needed which is able to detect a stamp in a scanned document and, in a second step, to differentiate whether it is authentic or forged.

Moreover, detection of stamps is important in other levels of document security too. For example in the scenario where there are invoices incoming from the same source, it is desirable to determine whether there are no outliers, for example a document with a missing stamp.

The difficulty of stamp extraction is that, in general, there is no template for stamps. It is a partially graphical and textual object that can be placed on any position in the document. The variations are in its shape and color, print quality or rotation and even two imprints of the same stamp can look very different.

We present a new approach to detect stamps of different colors and extract them properly even if they are overlapped with a signature or a text of another color. We do not focus on stamps of any particular shape so business stamps with addresses, official seals as well as decorative stamps can be extracted. Detection of black stamps still remains a challenge, but the method allows further extension for black stamps too. Good results have been achieved detecting stamps in documents containing other color objects such as logos and texts.

The proposed method consists of the following steps. First, the chromatic part of the image is separated from the approximately achromatic text and background. Then, color clustering is performed on it to obtain several cluster images, each containing just elements of the same color. The cluster images are cleaned from noise and segmented by XY-cut algorithm to extract candidate solutions. Finally, the candidate solutions are classified using the set of features described in Section IV.

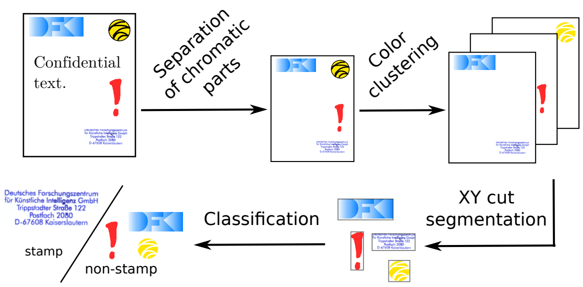


Figure 1. Diagram of the stamp detection algorithm.

In Section II, previous work on stamp detection is summarized. Section III describes the proposed method on segmentation of the document image and Section IV covers

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.227

feature extraction and classification. In Section V, the data set for evaluation is presented and the results are given in Section VI.

II. Related Work

The number of published methods on stamp segmentation is quite limited. As prior knowledge of the structure (shape or color) of the stamp is helpful to localize it in a document, previous research was focused on detection of stamps of a particular type.

Detection methods based on shape information require an outer frame to be present around the stamp. Chen et al. [2] detect seals on Chinese bank checks with a region-growing algorithm. They assume the seal to be the only object in the check to have an outer frame. The method proposed by Zhu [3] performs Hough Transform to search for elliptical/circular candidates for stamps. The author claims the method to be robust on degraded documents and successful even for stamps overlapping with text. However, due to the shape restriction is its usage limited.

In content-based image retrieval, the information carried in stamps (seals) contained in documents could be used for their indexing. Roy et al. [4] compute spatial feature descriptors from the positions of characters in the queried seal and apply Generalized Hough Transform to detect the seal in documents in the database. In this scenario, the template (user query) seal must be given.

A stamp is a plain-color object although some parts of the imprint might have different brightness due to an imperfect ink condition. Ueda [5] was one of the first authors to apply color analysis for extraction of signatures and seals from Japanese bank checks. He works with the RGB color histogram of pixel intensities and makes orthogonal projections to separate different clusters. The author assumes exactly 3 clusters to be in the document – a background, a seal and a signature – and such an approach is not suitable for generic document images. Cai and Mei [6] also present a simple approach based on color analysis dividing the RGB cube into 8 subspaces, assigning a label to each pixel and choosing just those pixels that belong to the red and blue subspace. With this approach, the color of the seal must be known beforehand. Soria-Frisch presents [7] a fuzzy integration method for combining color channels to segment stamps of one particular color.

Berger et al. [8], [9] are able to separate overlapping objects of very similar colors (e.g. a stamp and a signature) by means of SVM. However, small areas belonging to each object have to be chosen manually which makes their approach inapplicable to automatic segmentation.

In the latest method proposed by Forczmański and Frejlichowski [10] in 2010, the authors transform the document image into YC_bC_r space and work with C_b and C_r components which carry the information about chromaticity. Areas with high C_b or C_r intensities denote the presence of a stamp. Row and column projections are made on each of the two images separately to detect these areas. Simple features (size and width-to-height ratio) are applied to discard inappropriate candidate solutions. This approach is restricted to detection of red and blue stamps only.

Evidently, all the published methods have considerable limitations (on shape, color or background) and therefore a generic approach to stamp detection is proposed in this paper.

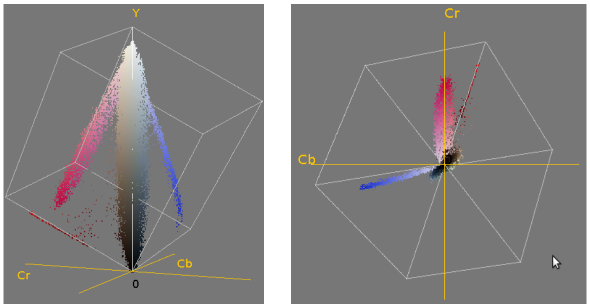
III. Image Segmentation

A stamp is treated as a single color object. This characteristic is considered to be invariant and it is utilized for segmentation of the document image. Only chromatic parts are extracted and clustered according to their color. The clusters are then partitioned to obtain candidate solutions.

A. Separation of Chromatic Pixels

RGB color model is not convenient for segmentation because of high correlation among the channels [11]. To segment color stamps, it is desirable to separate out the gray-level (achromatic) parts of the document image that correspond to the printed text and background. For this purpose, YC_bC_r color space has been chosen. In this model, Y is the luma component and C_b, C_r are chroma components meaning the blue and red difference [12].

$$\begin{aligned} Y &= 0.299 \cdot R + 0.587 \cdot G + 0.114 \cdot B \\ C_b &= -0.169 \cdot R - 0.331 \cdot G + 0.5 \cdot B \\ C_r &= 0.5 \cdot R - 0.419 \cdot G - 0.081 \cdot B. \end{aligned} \tag{1}$$



(a) (b)

Figure 2. Ink colors in the image form clusters of elongated shapes. A scatter plot in YC_bC_r color space in (a) and a projection of pixel intensities onto the C_bC_r plane in (b).

To separate out the pixels close to gray levels, a projection on C_bC_r plane is made and each pixel point is treated as a polar vector (r, θ), where $r = \sqrt{C_b^2 + C_r^2}$ and $\theta = atan2(C_b, C_r)$, $\theta \in [0, 2\pi)$. A threshold T is set and all vectors with $r > T$ are marked as chromatic and further used for color clustering.

B. Color Clustering in C_bC_r plane

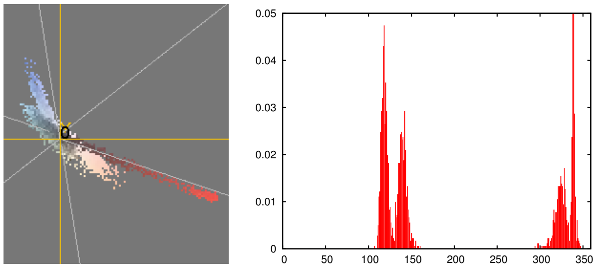
From pixel scatter plots of various document images one can learn that the clusters formed by inks have elongated shapes like clouds stretching from the white color cluster as it can be seen in Fig. 2(a). Projecting the pixel vectors onto C_bC_r plane (see Fig. 2(b)), the ink clusters stretch from the center of the coordinate system and they are quite narrow. The polar angle θ of each vector is discriminative for segmentation and the advantage of this property has been taken. It should be noted that only pixels labeled as chromatic in the previous step are further processed.

1) Polar coordinates of all pixel vectors in C_bC_r were already computed in the previous step. Only the polar angles θ_i will be further needed for clustering.
2) The angle values are quantized into 360 values and a histogram is constructed. Let us denote the bin of θ_i as θ'_i, an integer value from interval $[0, 360)$. An example is given in Fig. 3.
3) For two vectors $\vec{u_1} = (r_1, \theta_1)$, $\vec{u_2} = (r_2, \theta_2)$, the definition of their distance is:

$$d(\vec{u_1}, \vec{u_2}) = \begin{cases} |\theta'_2 - \theta'_1|, & \text{if } |\theta'_1 - \theta'_2| \leq 180 \\ 360 - |\theta'_2 - \theta'_1|, & \text{if } |\theta'_1 - \theta'_2| > 180. \end{cases}$$

4) The histogram is used to determine the number of clusters by an approximate numeration of peaks. The peak bins are also set as the initial cluster centers.
5) The k-means clustering algorithm is run in the histogram space with the above defined distance metrics.

Performing k-means clustering on histogram space accelerates the computations dramatically.



(a) (b)

Figure 3. The projection of image pixel intensities on C_bC_r plane is depicted in (a) and the corresponding histogram of quantized angle values is shown in (b). Four peaks in the histogram correspond to the four clusters in C_bC_r plane. The y-axis values is the percentage of amounts of pixels in the whole document.

C. Extraction of Candidate Solutions

The derived clusters are handled as separate binary masks of the original image. They are noisy because even the textual part of the document image contains chromatic pixels due to the light diffusion effect during scanning. However, the noise can be very efficiently removed by applying the morphological *opening* operation.

By nature, stamps are not connected objects – they consist of many small components that need to be grouped properly.

The resulting cleaned masks can be segmented by the well-known XY-cut algorithm [13]. The resulting rectangles are handled as bounding boxes of the candidate solutions. With this approach even multiple stamps can be extracted.

IV. Classification

To differentiate stamps from other color objects in documents such as logos, pictures or a color text, geometrical and color-related features are extracted from the candidates.

A. Geometrical Features

A simple geometrical feature is *width-to-height ratio*. To discard too long objects such as vertical or horizontal lines, it should not exceed 20 and be less than $1/3$.

Area of the bounding box and *pixel number* reflect the size of the candidate object. However, different scan resolutions and sizes of the document have to be taken into account. Therefore the feature values are not absolute but relative to the average size of connected components counted from the whole document image. For that purpose the image is first converted to grayscale and binarized by Sauvola's adaptive method [14] that is particularly effective for OCR. Connected components are labeled and their bounding boxes determined. Noisy small components are discarded and the others are ordered according to the area of their bounding boxes. Then the median size is selected.

Pixel density within the bounding box (BB) of the candidate is restricted:

$$T_1 < \frac{\text{segmented pixels}}{\text{all pixels in BB}} < T_2. \qquad (2)$$

According to our experiments, $T_1 = 0.03$ and $T_2 = 0.5$ are suitable values for the thresholds. The lower bound assures that noisy candidates with very sparse masks will be rejected and the upper bound helps distinguishing stamp candidates from plain-color parts of logos.

A strong feature differentiating mainly business stamps with written names and addresses from random color text is *rotation*. It has been uncovered in a series of experiments held while gathering the data set that people tend to imprint stamps with a skew, though sometimes hardly perceptible by an untrained eye. The angle between average skew of the page and stamp text lines is normally greater than $0.4°$ while standard deviation of page text lines is around 0.2.

To compute the difference, skew of the page must be determined first. For this purpose, a skew detection method by Breuel [15] has been adopted. It performs text-line extraction by modeling the text-lines. A quality function is used to measure how a given set of points matches the text-line model. To maximize the number of bounding boxes matching the model and to minimize the distance of each reference point from the baseline, the RAST algorithm and branch-and-bound search are applied.

B. Color-related Features

Standard deviation of hue is a well-discriminative feature as hue remains similar throughout the imprint even if it was impressed irregularly with some regions being darker or brighter. Therefore, its standard deviation should be low. The circular property of hue must be taken into consideration and to compute the mean, all values must be treated relatively to a reference value (e.g. mode). The following restriction delimits the stamp values:

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\Delta x_i)^2} < T_3, \tag{3}$$

$$\Delta x_i = \begin{cases} |x_i - \bar{x}|, & \text{if } |x_i - \bar{x}| \leq 180 \\ 360 - |x_i - \bar{x}|, & \text{if } |x_i - \bar{x}| > 180, \end{cases} \tag{4}$$

where N is the number of pixels of the candidate, x_i are pixel hues in degrees, $\bar{x}$ is the mean and T_3 is a threshold set to 9.0 according to our experiments.

The proposed color clustering method allows for an efficient segmentation of color objects and enables stamps being overlapped with text of different colors to be extracted properly. However, due to this approach multi-colored objects are split and some context is lost. The presence of multi-colored objects close around the candidate solution often indicates a logo so it is a valuable information. That is why the mask of the candidate solution is dilated and the XOR operation is performed on the original and dilated mask images. The result is used to mask out the pixels in the close neighborhood of the candidate solution. By the same method as described in Section III-A, only chromatic pixels are selected and in case that their amount exceeds a minimum proportion, the candidate is rejected. The minimum must be set high enough not to reject stamps overlapped with signatures or texts.

V. Evaluation

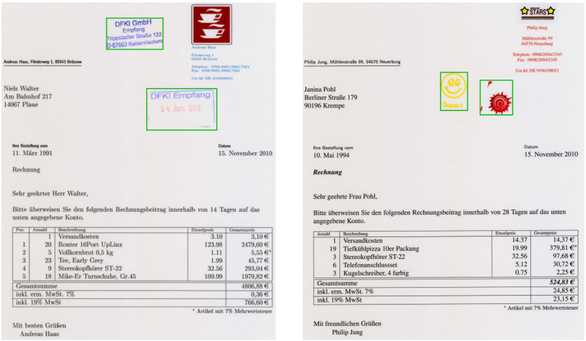
To the best of our knowledge there is no data set of color document images with stamps available to public for the purpose of evaluation. Tobacco800 data set is at hand but it just contains binary images with very few stamps that are not annotated. For that reason, a new data set containing 400 scanned document images has been collected and made public[1] encouraging other researchers to compare their results. The documents are automatically generated invoices that were printed, stamped and scanned with 600 dpi resolution. They include color logos and color texts which makes the evaluation results more realistic. There are stamps of many different shapes and colors including black ones in the data set, sometimes the stamps are overlapped with signatures or a text. In some documents there are multiple stamps or none at all. The groundtruth consists of binary images with masks of the stamp strokes which allows for accurate pixel-wise evaluation.

In Tab. I, the statistical information on the data set is summarized and in Fig. 4 some sample documents are given.

The evaluation was performed on the data set excluding documents with black stamps, which amounts totally 320 images. In 52 of these documents, the stamp is overlapped with other objects.

Documents with	Number of documents	Percentage of all
one stamp	285	71%
multiple stamps	56	14%
no stamp	59	15%
color stamp(s)	287	72%
black stamp(s)	80	20%
overlapped stamps	55	14%

Table I
STATISTICAL INFORMATION ON THE DATA SET.



DFKI GmbH
Empfang
Trippstadter Straße 122
D-67663 Kaiserslautern

Andreas Haas
Fliederweg 1
05543 Brüssow
Telephon: 0760-9985/8363/7264
Fax: 0760-9985/8363/7264
Ust-Id: DE 2195690918

Andreas Haas, Fliederweg 1, 05543 Brüssow

Niels Walter
Am Bahnhof 217
14067 Plaue

DFKI Empfang
04. Jan. 2011

Ihre Bestellung vom
11. März 1991

Datum
15. November 2010

Rechnung

Sehr geehrter Herr Walter,

Bitte überweisen Sie den folgenden Rechnungsbeitrag innerhalb von 14 Tagen auf das unten angegebene Konto.

Pos.	Anzahl	Beschreibung	Einzelpreis	Gesamtpreis
	1	Versandkosten	3.10	3,10 €
1	20	Router 16Port UpLinx	123.98	2479,60 €
2	5	Vollkornbrot 0,5 kg	1.11	5,55 €*
3	23	Tee, Early Grey	1.99	45,77 €
4	9	Stereokopfhörer ST-22	32.56	293,04 €
5	18	Mike-Er Turnschuhe, Gr.45	109.99	1979,82 €
Gesamtsumme				4806,88 €
inkl. erm. MwSt. 7%				0,36 €
inkl. 19% MwSt				766,60 €

* Artikel mit 7% Mehrwertsteuer

Mit besten Grüßen
Andreas Haas

(a)

STARS

Philip Jung

Mühlenstraße 99
04570 Neuerburg

Telephon: 0908/2694/2345
Fax: 0908/2694/2345

Ust-Id: DE 9196198021

Philip Jung, Mühlenstraße 99, 04570 Neuerburg

Janina Pohl
Berliner Straße 179
90196 Krempe

Super!

Ihre Bestellung vom
10. Mai 1994

Datum
15. November 2010

Rechnung

Sehr geehrte Frau Pohl,

Bitte überweisen Sie den folgenden Rechnungsbeitrag innerhalb von 28 Tagen auf das unten angegebene Konto.

Anzahl	Beschreibung	Einzelpreis	Gesamtpreis
1	Versandkosten	14.37	14,37 €
19	Tiefkühlpizza 10er Packung	19.99	379,81 €*
3	Stereokopfhörer ST-22	32.56	97,68 €
6	Telefonanschlussset	5.12	30,72 €
3	Kugelschreiber, 4 farbig	0.75	2,25 €
Gesamtsumme			***524,83 €***
inkl. erm. MwSt. 7%			24,85 €
inkl. 19% MwSt			23,15 €

* Artikel mit 7% Mehrwertsteuer

Mit freundlichen Grüßen
Philip Jung

(b)

Figure 4. Sample documents from the data set. Stamps detected by our algorithm are marked with green bounding boxes (for illustration).

Though the data set was scanned at resolution of 600 dpi which was the highest resolution available for the scanner, the experiments showed that the method has sufficiently good results even for low scan resolutions which are commonly used. The evaluation was performed on downscaled images with resolution corresponding to 300 and 200 dpi, which are scan resolutions normally used in offices.

VI. Results

The results are given in terms of pixel accuracy. *Recall* is the proportion of correctly detected stamp pixels to all stamp pixels in the image while *precision* is the proportion of correctly detected stamp pixels to all detected pixels.

As it can be observed from Tab. II, the results for 200 dpi scans are nearly the same as for 300 dpi so we can claim the method to be successful even on low resolution images.

The most challenging is the correct segmentation of stamps overlapped with other objects. A separate evaluation on 52 overlapped stamps was performed with recall of 69% and precision of 68%. An example is given in Fig. 5. Errors were caused by an improper segmentation of stamps overlapped with an object of a similar color. Further, there

[1] The data set is available at http://madm.dfki.de/downloads-ds-staver.

are a few cases in the data set that the stamps were not clean before coloring and impressing, so they produced a two-color imprint. Besides, a few logos are tricky and have similar features as stamps.

	Recall	**Precision**
300 dpi	83.4%	83.8%
200 dpi	82.7%	82.8%

Table II
THE RESULTS OF EVALUATION.

VII. CONCLUSION

A new approach for detection of stamps in color document images was presented in this paper. It involves conversion into YC_bC_r color space and extraction of chromatic pixels. Ink color clusters in YC_bC_r have elongated shapes and this property is exploited for image segmentation by color clustering. Candidate solutions are obtained and classified based on several geometrical and color-related features.

Stamps of any shape and arbitrary colors except for black can be detected by the algorithm. Nevertheless, the method is extensible for detection of black stamps too. A more accurate segmentation algorithm would have to be applied to the achromatic part of the image, the same geometrical features could be used and new features e.g. concerning the printing differences would have to be added.

A thorough analysis of previous work has been done and we claim the here presented approach to stamp detection to be the most generic one so far.

A new data set of 400 documents with stamps of different shapes and colors has been collected and made public. The method was evaluated on it with recall as well as precision of 83%. Evaluation on documents with low resolution corresponding to 200 dpi scans lead to the same results.

As a part of ongoing work, preliminary experiments suggest that the method can also distinguish between authentic stamps and some photocopies.

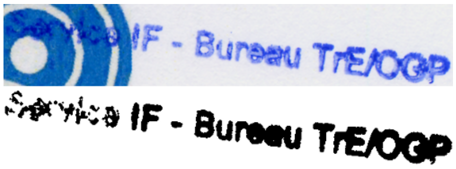


Figure 5. A correctly detected and segmented stamp from the data set.

ACKNOWLEDGMENT

Part of this work has been supported by the Danish Council for Independent Research – Technology and Production Sciences (FTP), grant 10-081972. The authors would also like to thank to Assoc. Prof. Adam Herout from Dep. of Computer Graphics and Multimedia in Faculty of Information Technology, Brno University of Technology, for his inspiring comments and assistance.

REFERENCES

[1] J. van Beusekom, "Optical document security in high volume office environments," PhD, University of Kaiserslautern, Germany, 2010.

[2] L. Chen, T. Liu, J. Chen, J. Zhu, J. Deng, and S. Ma, "Location algorithm for seal imprints on Chinese bank-checks based on region growing," *Optoelectronics Letters*, vol. 2, pp. 155–157, 2006.

[3] G. Zhu, S. Jaeger, and D. Doermann, "A robust stamp detection framework on degraded documents," in *Proc. of SPIE Document Recognition and Retrieval XIII*, 2006, pp. 1–9.

[4] P. P. Roy, U. Pal, and J. Lladós, "Seal object detection in document images using GHT of local component shapes," in *Proc. of the 25th ACM Symposium on Applied Computing*, 2010, pp. 23–27.

[5] K. Ueda, "Extraction of signature and seal imprint from bankchecks by using color information," in *Proc. of the 3rd Int. Conf. on Doc. Analysis and Recognition*, 1995, pp. 665–.

[6] L. Cai and L. Mei, "A robust registration and detection method for color seal verification," in *Proc. of the 1st Int. Conf. on Intelligent Computing*, 2005, pp. 97–106.

[7] A. Soria-Frisch, "The fuzzy integral for color seal segmentation on document images," in *Proc. of the 10th Int. Conf. on Image Processing*, 2003, pp. 157–160.

[8] C. E. Berger, J. A. Koeijer, W. Glas, and H. T. Madhuizen, "Color separation in forensic image processing," *Journal of Forensic Sciences*, vol. 51, pp. 100–102, 2006.

[9] C. E. Berger and C. J. Veenman, "Color deconvolution and support vector machines," in *Proc. of the 3rd Int. Workshop on Computational Forensics*, 2009, pp. 174–180.

[10] P. Forczmański and D. Frejlichowski, "General shape analysis applied to stamps retrieval from scanned documents," in *Proc. of the 14th Int. Conf. on Artificial intelligence: Methodology, Systems, and Applications*, 2010, pp. 251–260.

[11] H. D. Cheng, X. H. Jiang, Y. Sun, and J. Wang, "Color image segmentation: advances and prospects," *Pattern Recognition*, vol. 34, pp. 2259–2281, 2001.

[12] E. Hamilton, "JPEG file interchangable format," available from http://www.jpeg.org/public/jfif.pdf, C-Cube Microsystems, California, Tech. Rep., 1992.

[13] M. Krishnamoorthy, G. Nagy, S. Seth, and M. Viswanathan, "Syntactic segmentation and labeling of digitized pages from technical journals," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, pp. 737–747, 1993.

[14] J. Sauvola and M. Pietikainen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33, pp. 225–236, 2000.

[15] T. M. Breuel, "Robust least square baseline finding using a branch and bound algorithm," in *Proc. of SPIE Document Recognition and Retrieval IX*, 2002, pp. 20–27.