An improved scene text extraction method using Conditional Random Field and Optical Character Recognition

Hongwei Zhang [a] , Changsong Liu [a], Cheng Yang [a], Xiaoqing Ding [a], KongQiao Wang [b]
[a] State Key Laboratory of Intelligent Technology and Systems
Department of Electronic Engineering, Tsinghua University, Beijing, P.R.China
{zhanghongwei, lcs, yangcheng, dxq}@ocrserv.ee.tsinghua.edu.cn
[b] Nokia Research Center Beijing, BDA, 100176, P.R.China

Abstract—**Over the past few years, research on scene text extraction has developed rapidly. Recently, condition random field (CRF) has been used to give connected components (CCs) 'text' or 'non-text' labels. However, a burning issue in CRF model comes from multiple text lines extraction. In this paper, we propose a two-step iterative CRF algorithm with a Belief Propagation inference and an OCR filtering stage. Two kinds of neighborhood relationship graph are used in the respective iterations for extracting multiple text lines. Furthermore, OCR confidence is used as an indicator for identifying the text regions, while a traditional OCR filter module only considered the recognition results. The first CRF iteration aims at finding certain text CCs, especially in multiple text lines, and sending uncertain CCs to the second iteration. The second iteration gives second chance for the uncertain CCs and filter false alarm CCs with the help of OCR. Experiments based on the public dataset of ICDAR 2005 prove that the proposed method is comparative with the existing algorithms.**

Keywords-CRF;OCR;BP; Scene text extraction

I. INTRODUCTION

Currently, with the development of low-priced, portable and high performance digital imaging device, more and more people enjoy capturing interesting scene text with digital camera or mobile phone and sharing them on the internet. This requires a method of obtaining text information from nature scene image automatically, instead of typing-in manually. As one of the clear applications of pattern recognition, OCR (Optical Character Recognition) has been viewed as a solved problem, while text extraction turns to be a far more challenging problem.

Scene text extraction has long been known as a hard problem because of the difficulties in text segmentation from the varied and complicated backgrounds. Traditional algorithms tend to classify individual image regions into text or non-text regions using texture features or shapes features. Although some prior methods perform promisingly well in images with clear text lines, their performances drop dramatically when detecting texts in complex backgrounds. Actually, text lines or words can be modeled as multi-part objects and contextual connections between neighboring objects are of great importance in text line extraction.

Markov Random Field Models (MRF) theory is a tool to encode contextual constraints into the prior probability while Conditional Random Field (CRF) is an improvement method of MRF (Markov Random Field) model. MRF and CRF based approaches have been successful in modeling low level vision problems such as image restoration, segmentation [4], etc. Investigation of MRF and CRF modeling in high level vision such as object matching and scene text extraction, which is more challenging, begins only recently.

D.Q. Zhang [5] proposed a parts-based approach for 3D scene text detection using a three-order MRF model. A burning issue in MRF or CRF model comes from multiple text lines extraction, for the negative constraint produced by the cross-text-line clique usually leads to a terrible miss of detection. D.Q Zhang fixed this problem by modifying the potential function such that it only has positive constraint effect, which brought additional noises [5]. However, CRF is more appropriate for text extraction: in a CRF model, the potential is a function of all the observation data as well as the label, while only one observation is considered in MRF model [8]. In a recent work, Y.F Pan [6] presented a CRF model for component analysis using the text confidence as a unary feature. However, as human beings recognize a scene text not only by its style but also by its meanings, an OCR filtering stage is crucial, as a terminate step, to be used for discarding false positive text regions. A recent work [7] used OCR reading result to filter regions beyond recognition. Nevertheless, cases are that text-like background regions are recognized as text characters with a low confidence.

In this paper, we propose a two-step iterative CRF algorithm with a BP inference and an OCR filtering stage. The proposed method utilizes OCR confidence as an indicator for identifying the text regions. Furthermore, two kinds of neighborhood relationship graph (NRG) are used in the respective iterations for multiple text lines extraction. The two-step iteration works as follows: in the first CRF iteration, connected components (CCs) with high OCR confidence are identified as text regions, while uncertain CCs are reclassified in a second iteration process. Experiments show that the proposed method performs more effective in reserving text regions and discarding false positive background than the existing algorithms.

II. SYSTERM OVERVIEW

The proposed method includes two iterative processes, each of which can be divided into three stages: (1) CC extraction; (2) CRF classification, (3) CC filtering. Fig1 shows the pipeline of this algorithm.

Two kinds of NRG are built using different rules. The strict NRG guarantees that each pair of neighbor CCs has

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.148

overlapped horizontal projection which prevents cross-text-line clique. The relaxed NRG contains all pairs of node which have closed related pixels so that background CCs can be connected as much as possible.

In the first iteration, the strict NRG is applied in CRF and CCs are labeled as 'text' or 'non-text'. CCs labeled as 'text' are grouped into rectangle regions and sent into an auto-segmented OCR module. The OCR module gives each rectangle region a confidence (OCR confidence) of being a text line. Regions with high OCR confidence are identified as text lines and corresponding CCs identified as text CCs (Strict OCR). Others are called uncertain CCs.

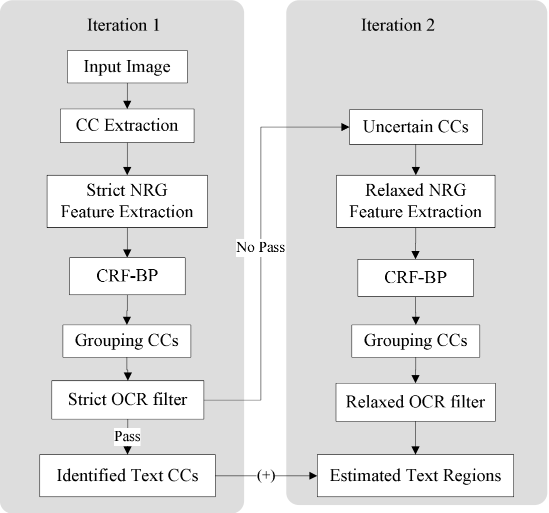

Figure 1. Flowchart of the proposed method

In the second iterative step, CRF utilizes relaxed NRG without the identified text CCs. OCR confidence threshold is set to a lower level (Relaxed OCR) so as to preserve text CCs as possible. The estimated text regions are made up of the result in the second iteration and the identified text regions.

III. Connected Component Extraction

Scene text usually enjoys a distinguishable color and brightness so it stands out from the background. What's more, characters in a specific text line tend to have a similar color. Under this assumption, text components can be separated from backgrounds based on different color properties they possess.

A. *Image Binarization*

A double threshold Niblack binarization algorithm [9] is applied on the grayscale result of a scene image.

$$L(x,y)=\begin{cases} 0 & I(x,y)-T<=-D \\ 255 & I(x,y)-T>D \\ 128 & others \end{cases} \qquad (1)$$

Where *"T"* is the average grayscale value within a rectangle window, which is centered on position *(x,y)*. *"D"* is a margin value and *"L"* is the quantization result for a specific pixel.

Accordingly, two binary images are created to pick up bright text and dark texts, respectively. The two images are processed separately and the results are merged at the end.

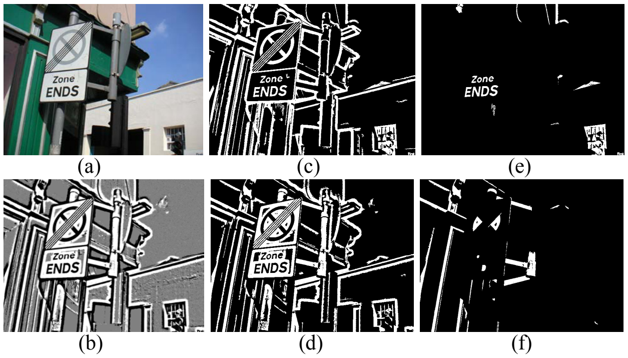

Figure 2. Example of the CC extraction: (a) origin image. (b) binarized image. (c) CCs in the black layer. (d) CCs in the white layer.(e) filtered CCs in black layer. (f) filtered CCs in white layer.

B. *CC Filtering*

As scene text within one line usually has similar size, we abandon CCs with exaggerated size or aspect ratio. What's more, isolated CCs without analogues neighbors are carefully rejected. Figure 2 shows an example of CC extraction process.

IV. Conditional Random Field Modeling

A. *Neighborhood relationship graph*

We use r_i to represent an individual CC. n_i is the total number of pixels in r_i. The rules for building NRG are as follows:

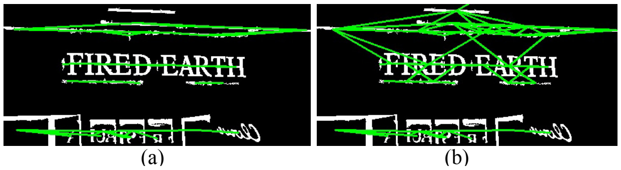

Figure 3. Example of NRG (a) Strict NRG (b) Relaxed NRG

- *Rules for Relaxed NRG:*

$$\left.\begin{array}{l} D(r_i,r_j)<min(max(w_1,h_1),max(w_2,h_2)) \\ A(r_i,r_j)<\pi/4 \end{array}\right\}\Rightarrow r_i=N^r{}_j, r_j=N^r{}_i \qquad (2)$$

Where

$$D(r_i,r_j)=\min_{p\in r_i,q\in r_j}\sqrt{(p_x-q_x)^2+(p_y-q_y)^2}$$

$$A(r_i,r_j)=\arctan\left(\frac{c^j{}_y-c^i{}_y}{c^i{}_x-c^j{}_x}\right)$$

$$(c^i{}_x,c^i{}_y)=\left(\frac{1}{n_i}\sum_{p\in r_i}p_x,\frac{1}{n_i}\sum_{p\in r_i}p_y\right)$$

- *Rules for Strict NRG:*

$$\left.\begin{array}{l} Rules\ for\ relaxed\ NRG \\ overlapped\ vertical\ projection \end{array}\right\}\Rightarrow r_i=N^s{}_j, r_j=N^s{}_i \qquad (3)$$

Figure 3 displays an example of the strict NRG and relaxed NRG.

B. Feature

CRF Feature aims at effectively distinguish text CCs from background noise. The following unary features are used in the proposed system:

- Size: if the CC is too big or too small, it should be discarded.
- Average gradient: Text CCs usually have obvious boundary. Sobel operator is used here for gradient image calculation.
- Aspect ratio: long bar-shaped CCs can be removed by this rule.

$$AspectRatio = \min(h/w, w/h) \tag{4}$$

Where *h* and *w* are the height and width of a CC.

As text characters generally appear together with other characters of similar properties, the following binary features are considered:

- Nearest distance: it is the same as $D(r_i, r_j)$ which has been defined forehead.
- Height ratio:

$$Hdis = \min(h_1/h_2, h_1/h_2) \tag{5}$$

Where h_1 or h_2 indicates the height of the external rectangle of CC.

- RGB Color distance: average of the color difference in the three layers.

$$D = \frac{1}{3}(|M_{R1}-M_{R2}| + |M_{G1}-M_{G2}| + |M_{B1}-M_{B2}|) \tag{6}$$

Where M_{R1} is the 1st CC's mean color on the red channel, and so on.

- Alignment: Angle of the connection line between the two CCs' center points.
- Stroke width variation: we devised a way to extract stoke width using image erode algorithm.

We erode the CC gradually until the whole CC is licked up. Meanwhile, the curve of the remaining area (in pixels) is drawn. The maximum point of the second derivative of the changing curves is considered as the stroke width of current CC.

C. CRF model description

The neighborhood relationship can be regarded as an undirected graph model and each CC equivalents to a node of it.

As for text extraction problem, classifying CCs into text or non text categories is realized by labeling them with 1 and 0, where 1 represents text and vice versa. A CRF model solves the problem of finding an optimized label by minimizing an energy function.

The label set $F = \{f_{i,...,}f_n\}$ is said to be a CRF conditioned on observations $R = \{r_i, ..., r_m\}$, if every f_i satisfies the Markovianity [8].

$$p(f_i \mid R, f_{S-\{i\}}) = P(f_i \mid R, f_{N_i}) \tag{7}$$

According to the Markov-Gibbs equivalence [8], we have:

$$p(F \mid R) = \frac{1}{Z}\exp(-\frac{1}{T}E(F \mid R)) \tag{8}$$

Where Z is the partition function and *E(F|R)* the energy function[3]. If only up to pairwise clique potentials are nonzero, the energy function has the form [3]

$$E(F \mid R) = \sum_{i\in S} D(f_i, r_i) + \alpha_{ij}\sum_{i\in S}\sum_{j\in N_i} W(f_i, f_j, r_i, r_j) \tag{9}$$

where r_i are features of the observed nodes, $D(f_i, r_i)$ and $W(f_i, f_j, r_i, r_j)$ are potential functions which has to be estimated in learning process.

D. Max-Product Belief Propagation

The max-product belief propagation algorithm is effective on optimizing MAP solution of CRF model problems by passing messages around the CC graph defined by the neighborhood relationship [9].

Let $m^t_{p\to q}$ be the message that node p sends to a neighboring node q at iteration t. Originally, all entries in $m^0_{p\to q}$ are initialized to zero. At each iteration step, new messages are passed within neighboring nodes; the message value is computed as follows:

$$m^t_{i\to j}(f_j) = min(W(f_i, f_j, r_i, r_j) + D(f_i, r_i) + \sum_{k\in N_i k\neq j} m^{t-1}_{k\to i}(f_i)) \tag{10}$$

The belief vector, indicating possibilities belong to each category (text, non-text), is calculated after T iterations in the following way:

$$b_j(f_j) = D(f_i, r_i) + \sum_{i\in N_j} m^T_{i\to j}(f_j) \tag{11}$$

Where $b_j(f_j)$ represents the belief that node j is labeled with f_j.

Finally, the label f_j^* that minimizes $b_j(f_j)$ individually at each node is selected.

E. Learning prior probabilities

The definition of potential function is of crucial importance for text extraction as it directly determines the expression of the energy function.

Theoretically, the potential function can be easily devised using the conditional probability which is available by statistics analysis.

We use y_i and y_{ij} to represent the unary and binary feature vectors. Both unary and binary conditional probability is assumed to be mixture of Gaussians.

The prior probability is learned from a set of scene images with or without text. We extract features from each pair of CC and prior probabilities are available accordingly.

V. Post Processing

A. Grouping CCs

CRF gives each CC a property label. However, the output of each text extraction algorithm is a set of rectangles designating bounding boxes for detected words. To group single CCs into text lines, NRG can be taken into account. A more restrictive condition is added, as well as the strict rule which is mentioned in section IV, for text line formation.

$$\begin{cases} C1 = \max(\left|r(r_i) - l(r_j)\right|, \left|l(r_i) - r(r_j)\right|) < 3 * \min(h(r_i), h(r_j)) \\ C2 = \max(\left|s(r_i) - s(r_j)\right|, \left|e(r_i) - e(r_j)\right|) < 2 * \min(h(r_i), h(r_j)) \\ C3 = \left|h(r_i) - h(r_j)\right| < \min(h(r_i), h(r_j)) \end{cases} \tag{12}$$

The forehead conditions are effective for occasional links between text lines and noise CCs which is unfortunately labeled as text in CRF module.

B. Region Filtering Stage

TH-OCR engine, which achieves superior recognition accuracy, is used in the proposed system for CC filtering.

We use distance information outputted by OCR to inversely represent the recognition confident, smaller is better. The absolute majority of text regions can be recognized correctly and gain a small distance except characters in a fancy style. Non-text regions are frequently recognized as uncommon symbols or gain a large distance in case it is recognized as letters or Arabic numerals. It is reliable for us to discard regions with low OCR confidence.

According to the language environment of involved scene images, it is reasonable of replacing the distance of a fallacious recognition result with a large value. For example, we set the uncommon symbols with a punishment distance of 200 (P-Dis) and the classifying threshold(C-Thre) 100. After punishment, regions with an average OCR distance (P-Ave) of lower than C-Thre are reserved. Figure 4 shows an example of the two-step iterative CRF algorithm with an OCR filtering stage.

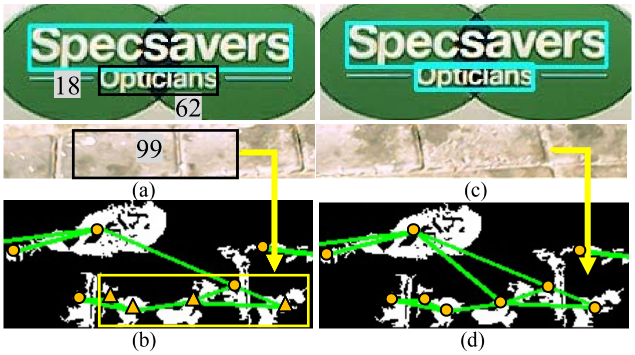

Figure 4. Example of the two-step iterative CRF algorithm. P-Ave is labeled next to corresponding region. (a).Text extraction result in the 1st iteration. The region in black rectangle is uncertain area when filtered with a strict OCR (P-Dis is set to 400 and C-Thre 50). (b).CRF processing (on strict NRG) result of the uncertain region in (a), text CCs are labeled with trangles and non-text ones with circles. (c).Text extraction result in the 2nd iteration using relaxed OCR (P-Dis is set to 200 and C-Thre 100). (d).CRF processing (on relaxed NRG) result of the uncertain region in (a).

Overlapped rectangles occasionally appear within the same layer or different layers. They are estimated by the following factors: OCR confidence, Height Occupy Ratio and area.

TABLE I. EXAMPLES OF OCR RESULTS

Regions		
OCR	L i b r a r y	: (T A <
Distance	19 11 28 31 69 35 23	125 58 69 92 68
P-Dis	19 11 28 31 69 35 23	200 200 69 92 200
Average	31	82
P-Ave	31	152

VI. EXPERIMENT SHOW

A. Data sets and evaluation rule

In order to provide a baseline comparison, we ran our algorithm on the most common benchmark, ICDAR 2005 text locating competition data set which contains 260 training images and 253 test images, as our training and test sample. Text in the dataset differs in size, color, font, highlights and possibly arrange in irregular orientation.

We evaluate the performance of the proposed text extraction with the evaluation program which is downloaded from the ICDAR 2005 official website.

B. Training Potential Function

In the training process, each image in the training set are binarized and CCs on the white and black layers are filtered. 23497 candidate CCs are retained, including 19625 background CCs and 3872 text CCs. With the help of the ground truth published along with ICDAR 2005 dataset, we humanly labeled each of these CCs with text, non-text or ignoring label. We ignored abnormal text CCs such as those not fully extracted as a result of strong reflective effect and those so called non-text CCs which is actually contour of text CCs owing to our binarization method.

Potential functions are obtained by fitting the distribution of the features using second order Gaussian curve.

C. Testing

In CRF process, BP algorithm iterates 400 times. At the 1st CRF iteration, we preserve the identified text regions by limiting the P-Ave to 50 and punishing the uncommon recognizing results with a distance of 400. While in the 2nd iteration, C-Thre is set to 100 and the P-Dis is relaxed to 200.

TABLE II. EXPERIMENT RESULTS OF DIFFERENT STAGES

	Precision rate (%)	***Recall rate (%)***	***f (%)***
Binarization	41.5	52.6	41.3
1st-NOCR	51.9	57.3	49.9
1st-OCR	55.4	57.1	52.0
2nd-NOCR	53.5	57.1	50.9
2nd-OCR	56.7	56.9	52.7

As there are two iterations in the proposed algorithm, we vertically compared text extraction result with different stage: binarization, the 1st iteration without OCR (1st-NOCR), the 1st iteration with OCR (1st-OCR), the 2nd iteration without OCR (2nd-NOCR) and the 2nd iteration with OCR (2nd-OCR). Please refer to table 2 for detail.

To compare the performance of the proposed method with the international development, parallel comparison is made with the top participates of ICDAR 2005 competition and a recent CRF related work described in [6]. 'Proposed Method' in table 3 displays the performance of the proposed method under ICDAR 2005 evaluation program.

TABLE III. COMPARISION WITH OTHER METHODS

	Precision rate (%)	***Recall rate (%)***
1st–ICDAR2005	62	67
Y.F. Pan	71	67
Proposed Method	56.7	56.9
Selected Dataset	65.9	80.9

a. The test result are obtained on ICDAR 2005 competition dataset

One more issue requires explanation here. The proposed method aims at extracting text lines (see figure 5), instead of separated words as the ground truth of the ICDAR 2005 competition, which contributes to a reduction of precision and recall rate. Without considering this factor, the proposed method is evaluated again and the accuracy improved apparently. 'Selected Dataset' in table 3 represents the improved result on images containing one word in each line.

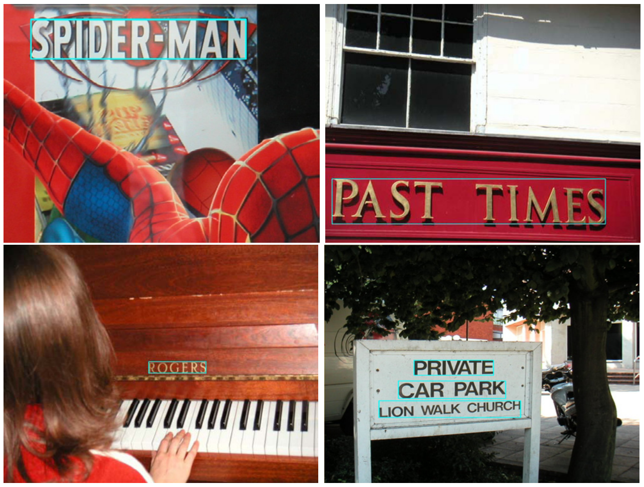

Figure 5. some text detection results on several images from the ICDAR 2005 test set, rectangels in green are detected text regions.

There are cases where texts are not detected, which are mainly due to single characters*, strong highlights, transparency of the text, size that is out of bounds, excessive blur, and curve baseline. Experiment proves that text region missing mainly occurs at the stage of image binarization, and the recall rate is essentially retained or even increased in the whole processing stages, please refer to table 2 for detail.

VII. CONCLUSIONS

In this paper, we present a two-step iterative CRF method for scene text extraction with OCR as a region filtering module. Regions with high OCR confidence in the first CRF iteration are reserved and uncertain regions are sent into CRF module with for a second classification. Experiments prove that in multiple text lines extraction, the negative constraint produced by the cross-text-line clique can be efficiently restricted by the strict and relaxed NGR. Furthermore, the OCR confidence is an effective parameter to tell a text region from background noises. There are several extensions for this work: the OCR filtering stage can be improved by a language mode and multi-scale framework can be added into the proposed system.

VIII. ACKNOWLEDGMENT

This work was supported by the National Basic Research Program of China (973 program) under Grant No. 2007CB311004 and the National Natural Science Foundation of China under Grant Nos. 60872086, 60772049.

REFERENCES

[1] Xiaolu S, Changsong L, Xiaoqing D, Yanming Z. Text line extraction in free style document. Proceedings of SPIE. 2009:72470L.

[2] Epshtein B. Detecting text in natural scenes with stroke width transform. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2010:2963-70.

[3] Liang J, Doermann D, Li H. Camera-based analysis of text and documents: A survey. International journal on document analysis and recognition. 2005;7(2-3):84-104.

[4] Cao H. Preprocessing of low-quality handwritten documents using markov random fields. IEEE Trans Pattern Anal Mach Intell. 2009;31(7):1184-94.

[5] Dong-Qing Zhang SC. In: Learning to detect scene text using a higher-order MRF with belief propagation. 27-02 June 2004; ; 2004. p. 101.

[6] Yi-Feng P, Xinwen H, Cheng-Lin L. Text localization in natural scene images based on conditional random field. 2009:6.

[7] Chen X, Yuille A. Detecting and reading text in natural scenes. PROCEEDINGS OF THE 2004 IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, VOL 2. 2004:366-73.

[8] Li SZ. Markov random field models in computer vision. Lecture Notes in Computer Science. 1994(801):361-.

[9] Felzenszwalb P, Huttenlocher D. Efficient belief propagation for early vision. International journal of computer vision. 2006;70(1):41-54.

[10] Lee S. Scene text extraction using image intensity and color information. Proceedings of the 2009 Chinese Conference on Pattern Recognition, CCPR 2009, and the 1st CJK Joint Workshop on Pattern Recognition, CJKPR. 2009:906-10.

[11] Lucas S, Panaretos A, Sosa L, Tang A, Wong S, Young R. ICDAR 2003 robust reading competitions: Entries, results, and future directions. International journal on document analysis and recognition. 2005;7(2-3):105-22.

[12] Lucas S. ICDAR 2005 text locating competition results. Eighth International Conference on Document Analysis and Recognition, Vols 1 and 2, Proceedings. 2005:80-4.

[13] W. Niblack, An Introduction to Digital Image Processing. Prentic Hall, Englewood Cliffs, NJ, 1986

* single CCs are labeled as non-text directly in the proposed system