Contribution to the Automatic Recognition of Business Documents

Djamel Gaceb, Frank Lebourgeois, Véronique Eglin, Hubert Emptoz

Contribution to the Automatic Recognition of Business Documents

Djamel GACEB *Frank LEBOURGEOIS* *Véronique EGLIN* *Hubert EMPTOZ*
LIRIS UMR 5205CNRS, INSA de Lyon 69621 Villeurbanne Cedex

djamel.gaceb1@insa-lyon.fr flebourg@rfv.insa-lyon.fr veronique.eglin@insa-lyon.fr hubert.emptoz@liris.cnrs.fr

Abstract

The automatic processing of paper documents and mails is a major challenge for all companies. Current recognition systems use modular architectures in which each stage of the process is independent. To improve the performances, it is necessary to reintroduce a co-operation between the different modules, for example by coupling the segmentation / recognition or zones of interests location / segmentation steps. In this context we propose a mixed approach for text localization and image segmentation which respects real time constraints. In the first part, we are going to present the state of the art in text location and thresholding in the images of postal addresses. In the second part, we will describe our method which simultaneously localize and segment text zones. The Location of text blocks obtained from a multiresolution approach on cumulated gradients computed directly from grey level images. The coupling of the two processes (text zones location and thresholding) allows to reduce simultaneously the computing time by processing only necessary parts of the image and by obtaining a better character segmentation for the OCR (Optical Character Recognition). We will present the results obtained from the implementation of our approach on an industrial line which daily processes several tons of documents from large companies.

Keywords: Text location, image segmentation, real time processing, business documents processing.

1. Introduction

The companies' mail automatic processing domain has several constraints:

- A very large variety of documents (text style, paper, texture, colour…).
- Real Time constraints
- An adaptation to the specificity of the image capture by linear cameras which requires optimized scan-line algorithms.
- Results obligation *(the system should be enough efficient to avoid an expensive manual processing of rejected documents).*

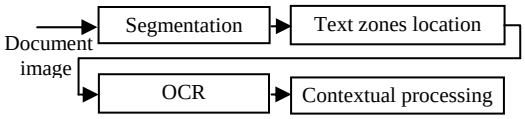
In addition to general constraints, our application has also particular constraints due to the specificity of the company demand:

- The images to be processed are divided into different categories corresponding to the mail families of company clients: handwritten internal mail (HIM), typing internal mail (TIM), forms (FMR), planus (PL) (PLN), bank credit cards(BC), A3 listing (LA3), A4 listing (LA4), address not found (NPAI), circulating bank checks (CCH). These images are very different in their size, orientation, background colour, text position and colour, characters size and text types (printed, printed matrix, handwritten…). The documents are processed by batches or arrive in bulk .
- We have approximately 1 second to achieve the image acquisition, the cleaning, the thresholding, the OCR analysis and the sorting decision-making. The images processing time is very limited and should not exceed 150 to 200 ms, for image acquisition, its thresholding and its text zones location.
- CCD camera current resolution is approximately 200dpi (10 inchs/2048 pixels) and can take only one image per document.
- The unrecognized documents are manually processed immediately. The failure recognition is generally explained by a dysfunction of the pre-processing stages and in particular the segmentation and location stages [4][5].

2. Comparing the existing methods

2.1. Revision of software architectures towards a cooperative approach

Existing software architectures in the mail sorting domain are essentially linear. The limits reached by the current vision systems are due to this data processing organization. The rejection and error rates of the industrial systems are high because of the independence of the processes engaged in the recognition.



This processes separation is adapted to the tasks repartition on several connected computers, but the failure of only one process step leads the system to reject or to make an interpretation error. Certain works refer already to more advanced architectures. [13] proposes a multi-agents system for data exchange and collaboration

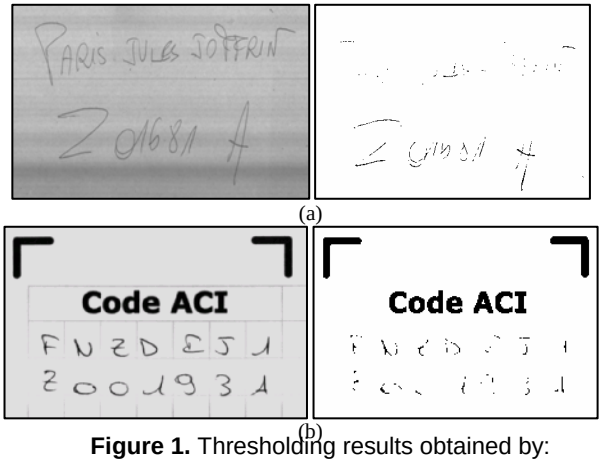
between the various acquisition and recognition modules. [21] has described a collaborative architecture of the various postal address and postal code recognition modules. [11] has described a probabilistic approach to combine the text location, segmentation and recognition. Finally, [29] described a words segmentation directed by recognition stage.

Our work aims to reduce rejection and error rates of the existing vision system by introducing a better cooperation between the various recognition stages while remaining within a real time process limits.

Thus, we will study a new nonlinear organization diagram of recognition process by introducing information feedback loops between various stages in order to make them cooperating. We will study possible information looping between the interest zones location, the text zones location, the segmentation stage, the OCR, the document pattern recognition and the document type classification… Among possible couplings, we propose in this article to start with a cooperation between the segmentation and text zones location stages. This cooperation should enable us to simultaneously save the processing times and improve the segmentation quality.

2.2. Comparing thresholding methods in the context of real time processing of business documents

We study the thresholding methods which are best suited to the specificity of the images we analyze, and which could satisfy the processing speed constraints and OCR adaptability. Automatic thresholding methods based on the histogram analysis such as the OTSU [18], Fisher [8], or based on entropy methods [1], are fast to calculate but they are ineffective on the slightly contrasted handwritten texts or on mixed machine-printed and handwritten documents(see Figure 1).



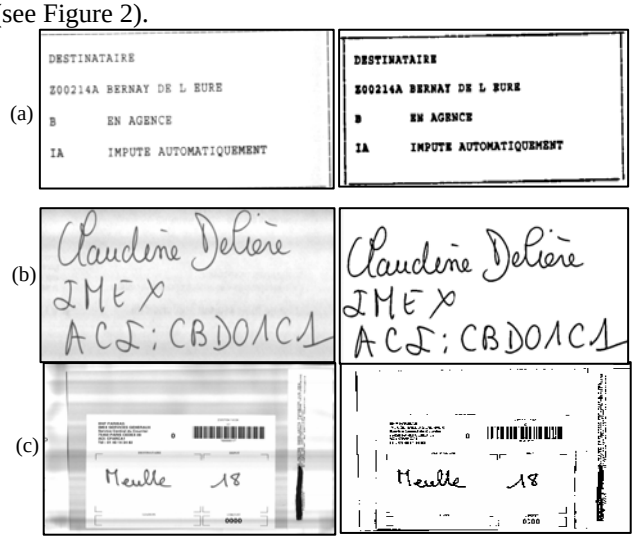
(a)

(b)

Figure 1. Thresholding results obtained by: (a) Otsu's method, (b) Fisher's method.

The adaptive threshold methods [16] [22] [27] calculate a threshold value for each image pixel by taking into account local information contained in its neighbourhood. If the window covers a zone of poorly-contrasted image, the detection threshold sensitivity is automatically increased. This adaptation to the local changes of the contrast explains the efficiency of these methods for the manuscripts images or the documents which use different ink colours.

The Niblack's method [16] has the disadvantage of increasing the detection sensitivity on the background zones thus highlighting the paper defects. The Sauvola [22] and Wolf [27] local methods correct these effects by limiting the detection sensitivity on the poorly-contrasted parts of the image, its give good results on the business documents segmentation of various origins containing both handwritten and machine-printed texts (see Figure 2).



(a)

(b)

(c)

Figure2: Thresholding by the Sauvola's local method , (a) CHC Image, window 7x7, (b) (CIM) Image, window 9x9, (c) 'Insert' Image, window 15x15.

The calculation costs of these methods are, however, very high since in order to calculate the associated threshold in each pixel, we should study the grey levels' distribution in a neighbourhood centred on this pixel. These times were calculated on a set of 9341 images of company internal mail using a pentium4 machine with a 2.22 GHz speed (Table1).

Thresholding Methods / Documents type	Niblack	Sauvola	Wolf	Fisher
CHC	4,44	4,47	4,38	*0,32*
NPAI	3,75	2,28	2,30	*0,29*
CB	4,34	3,46	4,30	*1,74*
LA3	4,38	4,46	4,32	*0,31*
LA4	2,42	2,43	2,42	*0,23*
FRM	3,59	3,50	4,33	*0,25*
PLN	4,53	4,53	4,45	*1,69*
HIM	4,50	4,61	4,39	*0,36*
TIM	1,30	1,28	1,29	*0,15*

Table 1. Summary of run speeds of the traditional thresholding methods (in seconds).

The adaptive (local) methods have the advantage to provide the best image segmentation even in difficult conditions. In the opposite these methods show several disadvantages:

- Prohibitive computing times which increase proportionally with the window size.
- Over segmentation of the paper defects and of the image noise on the empty zones of the image.
- Difficult processing of documents whose characters' size is different, in which the window size being fixed during all the processing.

None of the traditional (global or adaptive) methods fulfil all imposed conditions, namely a certain efficiency on all images for an imposed computing time.

2.3. Location of the text zones

Works on the address block location is grouped into several classes:

- Multi-resolution based methods.
- Aggregative methods by filtering.
- Bottom-up methods based on connected components.
- Images segmentation-based methods.
- Training-based methods.

The real-time constraints, the great variability of characters sizes and spaces between the words led several researchers to use the multi-resolution. The block address location by multi-resolution does not require a preliminary image thresholding. In addition, this location is based on a pyramidal construction enabling to reveal, on a suitable resolution scale, the structure of the bounding boxes lines [3]. Further, the pyramidal approach allows a top-down analysis to build an inclusion tree of segmented connected components on the various scales of resolution [23]. Other works use the traditional aggregative methods of the RLSA type [24] which are fast because they do not require an expensive connected components extraction.

These methods are, however sensitive to the documents slope and require a good orientation and a perfect text lines alignment. These aggregative approaches are not new. In fact, the first works where those of G. Nagy on texts location in the images to extract forms and mail physical structures in service of the great companies [15]. The computers of that time did not have the necessary computing power needed for advanced algorithms; he had the idea to use the progressive defocusing of the camera optics to make the image gradually fuzzy in which the characters become "spots" which agglomerate gradually to indicate the words, the lines and the text blocks.

The works on the addresses zones location based on a classification of the connected components are numerous and are not adapted to real time constraints.

Indeed, the capture of all connected components and the whole image "blind" thresholding are too expensive in terms of computing times [28]. Moreover these methods require a complex classification of connected components (CCs) according to their alignments and a rejection of the CCs which do not correspond to textual elements [19][20]. Finally these methods require a preliminary image thresholding.

The traditional segmentation-based methods such as Split & merge [25] make it possible to quickly locate the non-uniform regions of the image susceptible to contain the text. Other segmentation methods use also texture information with Gabor filters [6][2] or wavelets [12].

These methods locate at the same time the relevant image zones without capturing the connected components, but they differentiate the text areas of the non textual elements from their textures. However these interesting approaches are nevertheless very expensive in computing times.

Training-based location systems [7] [26] [14] appear difficult to implement in view of the large variety of documents that we have to process. Moreover certain researchers admit that the apprenticeship-based systems are less powerful than the systems whose rules have been adjusted manually to the problem put forward [17].

3. Our proposal

3.1. Thresholding/location Coupling

The separation between the two stages of thresholding and location of the texts simultaneously increases the computing time and leads to over-segmentations of the paper texture noise on the image background.

We have managed to optimize our thresholding method by applying adapted thresholds calculations near the text areas only. For that purpose we very quickly detect text areas in order to apply an adaptive thresholding like Sauvola along supposed text zones. We avoid in this way to binarize the empty zones which represent the major part of the image. It is then possible to apply purely local powerful methods to the zones containing the texts. The complexity of calculation of such an approach is not fixed. It fluctuates according to the place and of the distribution of the texts in the image. This approach will also allow to reduce the over-segmentation of the adaptive methods on the empty zones of the image.

3.2. Method used for text zones location

The location must be executed on grey levels image coming directly from the camera. The developed method must also, reduce as much as possible the number of false detections and adjust the zones in the text neighbourhood, so as not to waste time in segmenting the image background. We used a robust method, for quickly locating all text zones in a natural scene without particular lighting and without constraint during the image acquisition.

This process consists of agglomerating certain characteristic periodicities of the text lines which come from the luminous variations on characters contours or

generated by alternations between the lines or between the characters. These periodicities are calculated starting from pixels sequences with high gradients. To avoid filtering these points and introducing new thresholds, we carry out locally, in a neighbourhood V at each point (x_0,y_0), a simple magnitudes summation of the gradients normalized by number N of neighbourhood pixels $V(x_0,y_0)$.

$$G(x_0, y_0) = \frac{1}{N} \sum_{(x,y) \in V(x_0, y_0)} \frac{\partial f(x,y)}{\partial \vec{v}} \qquad (1)$$

This filter of "cumulated gradients", was initially developed for the text location in video frames [9], [27], [10].

However, its disadvantages towards our application oblige us to adapt it:

- The filter supposes that the text direction is a priori known. Indeed, the derivatives are calculated in the supposed text direction and summed in this same direction $\vec{v}$. To make filtering insensitive to the document image rotation, we will calculate and sum up the horizontal and vertical derivatives (2). We will use a rough but fast approximation for the calculation of the derivatives (3).
- The cost of the calculation of the summation in each image point in a neighborhood V is too high for our application. We are going to reduce this cost by carrying out the summation by blocs in multi-resolution. We divide the image into rectangular blocs of size dx × dy.
- For each bloc, we calculate the sum of the vertical and horizontal gradients (Figure 3).

$$J(x_0, y_0) = \frac{1}{dx\ dy} \sum_{i=1}^{dy} \sum_{j=1}^{dx} \left| \frac{\partial I(x_0 + j, y_0 + i)}{\partial x} \right| + \left| \frac{\partial I(x_0 + j, y_0 + i)}{\partial y} \right| \qquad (2)$$

and

$$\frac{\partial I}{\partial x}(u,v) = I(u-2,v) - I(u+2,v)$$
$$\frac{\partial I}{\partial y}(u,v) = I(u,v-2) - I(u,v+2) \qquad (3)$$

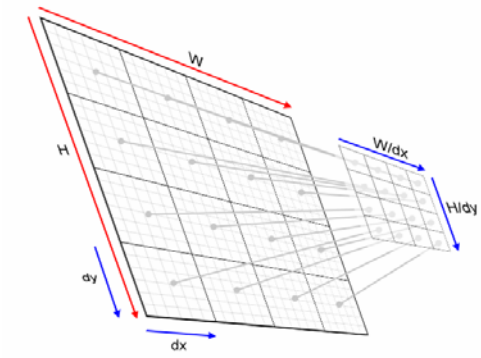


Figure 3: Image reduction and neighbourhood processing.

Figure 4: Thresholding steps

We obtain a reduced size image J (Figure 4) where the light zones represent the text areas. This bloc summation does not give the same results as the summation at each pixel.

We must carry out a morphological pre-processing on the reduced image J to obtain a filtering equivalent to that of the original algorithm. On this image, we apply consecutively the dilation d_1 times operator, e_1 times erosions, d_2 times dilations and e_2 times erosions (4).

$$K = E^{e2}(D^{d2}(E^{e1}(D^{d1}(J)))) \qquad (4)$$

These morphological transformations are used in order to:

- Agglomerate the text zones into blocs
- Take a rather sufficient margin on the writing to include the background containing relevant information (the texture and colour) necessary to obtain a better thresholding.

These mask parameters d_1, e_1, d_2, e_2 and the window size dx, dy are fixed arbitrarily for the moment. But the increase of dx and dy lead to a coarser and faster text detection whereas the increase of d1 and e1 detects better the texts zones agglomerated between them. Thus a study of the result stability on the various image types can lead to a satisfactory compromise.

- Text zones detection: $e_1 = d_1 = 2$ or 3
- Words detection: $e_1 = d_1 = 1$

The calculation over-cost of morphological operations is negligible because it is made on under-sampled image

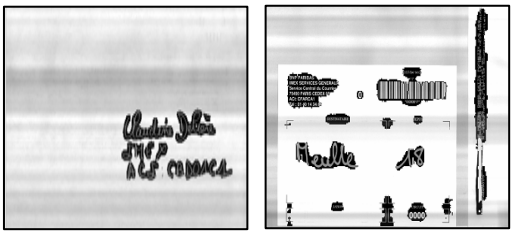


Figure 5. Zones to be binarized represented in inverted video.

3.3. Method used thresholding

We chose to use the Sauvola's method for its speed (Table 1) in the one hand, and for its performances in the other hand (the Wolf's method is specific to the video images and is not appropriate for our application). The times saved have enabled us to use a big size window (21 x 21) for Sauvola which makes it possible to obtain very good results on the printed or handwritten documents with very variable character sizes.

4. Results

We notice very well that by our method we reach run-speeds very similar to those of global binarisation (Table2) and much less important than local thresholdings. These times were calculated on a set of 9341 images of companies' internal mail.

Documents model	Half-locale & half- global method (HL&G)
CCH	0,56
NPAI	1,19
BC	1,42
LA3	0,51
LA4	0,27
FMR	0,68
PL	1,12
HIM	1,64
TIM	0,23

Table 2: Run-times of our algorithm (in seconds)

In addition to these run-speed improvements, we have also improved the recognition results (Table3). The results are an average of six days over six successive months knowing that the company process on average 29225 mails per day.

Mail model	Appreciation on OCR
LA3	+11%
LA4	+11%
NPAI	+26%
CCH	+2%
CB	+13%
TIM	+16%
HIM	+76%
PL	+20%

Table 3. Improvement of the recognition by our thresholding method (location/segmentation).

We have been able to obtain the best results, especially on the "handwritten envelopes" type of mail which contains many local variations:

Characters size variation: according to people writing style.

- Line thickness variation: according to pen, pencil or fluorescent used.
- Line colour variation: according to the colour of the pen used.
- Background variation due to various papers used for the internal envelopes (Kraft paper, plastic sheets).

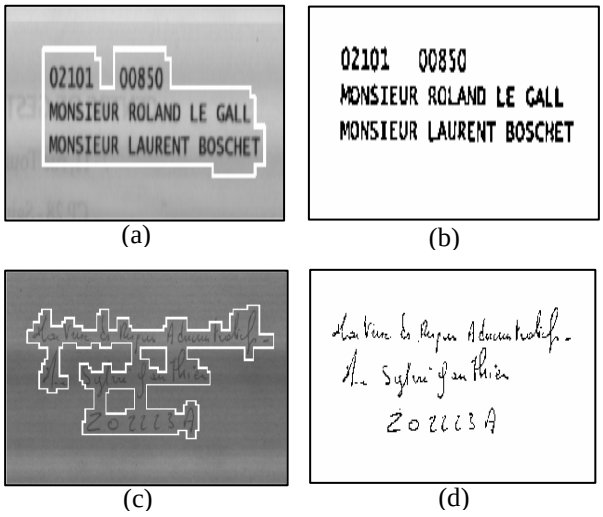


Figure 6: documents text-zone detection, (a) CID, (c) CIM, (b) and (d) thresholding results.

5. Conclusion and perspectives

By coupling text detection and image thresholding we could simultaneously reduce the calculating time and increase the segmentation quality. We could also improve and optimize the detection and recognition characters process.

The limited calculating time needs the choice of the more efficient method and its parameters (neighborhood window size, dilatations and erosions number) for a best compromise between binarisation quality and calculating times.

We could also, extend our combination of different recognition steps in order to guarantee the best cooperation and interaction between all the OCR system modules.

This work is granted by the CESA company (www.cesa.fr).

References

[1] A. S. Abutaleb, "Automatic thresholding of grey-level pictures using two-dimensional entropy", *computer vision graphics Image processing*, 1985, pp. 22-32.

[2] O. Deforges, D. Barba, "A fast multiresolution text-line and non text line structures extraction and discrimination scheme for document image analysis", *in proc of ICPR 94*, pp. 134-137.

[3] O. Deforges, C.Viard-Gaudin, D.Barba, "Gray-level Document Image Analysis", *2nd French-Korean Workshop, Man-Machine Handwritten Communication*, CNRS Ile de France, mai 1996, pp. 139-149.

[4] N. Gorski, and al., "A new A2iA bankcheck recognition system, Handwriting Analysis and Recognition", (Ref. No. 1998/440), *IEEE Third European Workshop on 14-15 July 1998*, pp.1-6.

[5] N. Gorski and al, "A2IA check reader", *ICDAR'99*, pp. 523-526

[6] A. K. Jain, Y. Chen, "Address block location using color and texture analysis, Computer Vision, Graphics and image processing", *image understanding*, sept 1994, 60 (2), pp.179-190.

[7] C. Jarousse, C. Viard-Gaudin, "Localisation du code postal par réseau de neurones sur bloc adresse manuscrit non contraint ", *CIFED'98*, Mai 1998, pp. 72-81.

[8] J. Fisher, S. Hinds, K. D'Amato, "A Rule-Based System for Document Image Segmentation ", *in proc. of the 10th Int'l Conf. Pattern Recognition, Atlantic City*, N.J., 1990, pp. 567-572.

[9] F. LeBourgeois, " Robust multifont OCR system from gray level images", *fourth ICDAR, International Conference on Document Analysis and Recognition*, Ulm, 1997, pp. 1-5.

[10] F. LeBourgeois, H. Emptoz , "Document Analysis in Gray Level and Typography Extraction Using Character Pattern Redundancies", *Int. Conf. On Doc. Analysis and Recognition ICDAR'99*, sep 20-22 1999, India, pp.177-180.

[11] Y. Lu and al., "An implementation of postal numerals segmentation and recognition system for Chinese business letters ", *ICDAR99*, pp. 725-728.

[12] D. Menoti and al., "Segmentation of postal envelopes for address block location : an approach based on feature selection in wavelet space", *In Proc. of ICDAR 03*, pp. 699-703.

[13] U. Miletzki, "Documents on the Move: DA&IR-Driven Mail Piece Processing Today and Tomorrow ", *in Proc. of DAS' 96*, pp. 547-563.

[14] U. Miletzki and al., "Continuous learning systems postal address readers with built-in learning capability ", *in proc. of ICDAR'99*, pp. 329-332.

[15] Nagy G.,"Preliminary investigation of techniques for automated reading of unformatted text ", *ACM, vol 11, n° 7, July 1968*, pp 480-487.

[16] W. Niblack, "An Introduction to Digital Image Processing ", *Englewood Cliffs, N.J.:Prentice Hall*, 1986, pp. 115–116.

[17] K. Nitz, "An Image-based mail facing and orientation system for enhanced postal automation ", *in proc. of ICDAR* '03, pp. 694-698.

[18] N. Otsu, "A threshold selection method from grey-level histogram", *IEEE trans system, man and cybernetics*, 1979, vol 9, pp. 62-66.

[19] J.C. Oriot, d. Barba, J.C. Salome, " Adress Block Locating Method Based On Transition Analysis Approach ", *Design And evaluation on flats objects, in proc. of ICDAR 91*, pp.665-673.

[20] J.C. Oriot, D. Barba, M. Gilloux, *Localisation du bloc adresse sur les objets postaux par une méthode de segmentation ascendante : évaluation et optimisation, Traitement du Signal*, 1995.

[21] S.N. Srihari, E.J.Kuebert, "Integration of hand-written address interpretation Technology into the United States Postal Service Remote Computer Reader System ", *in proc. of ICDAR 97, vol 2*, pp. 892-896.

[22] J. Sauvola, and al. "Adaptive Document Binarization ", *ICDAR'97, vol 1*, pp. 147–152.

[23] C. Viard-gaudin, D. Barba, "Localisation du bloc adresse par une approche multi-résolution ", *in proc. of ICDAR 91*, pp. 954-962.

[24] Wahl F, Wong K., Casey G., "Block segmentation and text extraction in mixed text/image documents", *Computer graphics and image processing*, 1982, n°20 p375-390.

[25] M. Wolf, H. Niemann, W. Schmidt, "Fast Address Block Location on Handwritten and Machine Printed Mail-piece Images ", *ICDAR 97,vol2*, pp. 753-757.

[26] H. Walischewski, "Learning regions of interest in postal automation", *ICDAR'99*, pp. 317-320.

[27] C. Wolf C, J.M. Jolion, F. Chassaing, "Text Localization, Enhancement and Binarization in Multimedia Documents", *In Proceedings of the International Conference on Pattern Recognition (ICPR)*, Quebec City, Canada, 2004, vol. 4, pp. 1037-1040.

[28] B. Yu, A. K. Jain and M. Mohiuddin, "Address Block Location on Complex Mail Pieces", *in proc. of ICDAR'97*, vol. 2, pp. 897-901.

[29] J. Zhou and al. "A feedback-based approach for segmenting handwritten legal amounts on bank cheques", *in proc. of ICDAR'01*, pp. 887-891.