# Word Separation in Handwritten Legal Amounts on Bank Cheques Based on Spatial Gap Distances

In Cheol Kim, Kyoung Min Kim, and Ching Y. Suen

Centre for Pattern Recognition and Machine Intelligence (CENPARMI),
Concordia University, 1550 de Maisonneuve Blvd. West, Suite GM606,
Montreal, H3G 1M8, Canada
{kiminc, kkm, suen}@cenparmi.concordia.ca

**Abstract.** This paper presents an efficient method of separating words in handwritten legal amounts on bank cheques based on the spatial gaps between connected components. Currently all typical existing gap measures suffer from poor performance due to the inherent problem of underestimation and overestimation. In order to decrease such burden, a modified version for each of those existing measures is explored. Also, a new method of combining three different types of distance measures based on 4-class clustering is proposed to reduce the errors generated by each measure. In experiments on real bank cheque database, the modified distance measures show about 3% of better separation rate than their original counterparts. In addition, by applying the combining method, further improvement in word separation was achieved.

## 1 Introduction

Automatic recognition of legal amounts on bank cheques has been intensively studied for many years in order to minimize manual efforts in bank cheque processing [1, 2]. However, due to the shape variability of the writers' unconstrained writing style, and horizontal overlapping, touching, and irregular gaps between connected components limited by space on bank cheques, legal amount recognition still remains as a challenging task. Separating a sentence into words and analyzing them is a typical approach to recognize a legal amount. In this case, precise extraction of the individual words is an essential step for achieving a reliable recognition performance.

In many related studies, measuring and sorting the spatial gaps between connected components using a specific distance measure has been employed as a typical approach to extract words. Seni and Cohen [3] proposed eight distance measures for extracting words from a handwritten text line. Among them, bounding box (BB) method that computes the horizontal distance between the bounding boxes of two adjacent connected components, and run-length/Euclidean with heuristics (RLEH) method that estimates the gap between two connected components using either the minimum run-length or the minimum Euclidean distance have shown superior performance. Mahadevan and Nagabushnam [4] proposed a convex hull (CH) method that approximates the gap between components by the distance between the convex hulls surrounding each component.

R. Orchard et al. (Eds.): IEA/AIE 2004, LNAI 3029, pp. 453-462, 2004.

The above three distance measures, BB, RLEH, and CH methods are simple and quite efficient but all suffer from the inherent problem of underestimation or overestimation. In order to reduce their drawbacks we first propose to modify each of these measures. In the case of BB method, the left and right boundaries of bounding boxes are adjusted appropriately to reduce the underestimation errors. Such concept of adjusting boundaries is also applied to the RLEH method to blunt its over-sensitivity to component shape caused by dealing with component contours directly. In CH method, a pair of convex hulls each of which respectively encloses the left and right part of a given connected component is newly introduced to avoid an overestimation problem. Next, a new method of integrating these three distance measures based on 4-class clustering technique is proposed to compensate effectively the errors in each measure, whereby further improvement in performance of word separation is expected.

In word separation experiments on a legal amount database extracted from real-life bank cheques, we demonstrate the effectiveness of the modified distance measures and combining method based on 4-class clustering by comparing their separation rates with those of the original distance measures.

## 2 Distance Measures for Gap Estimation

For a word separation task, we first employ three well-known distance measures, BB, RLEH, and CH method, and investigate their geometrical properties and drawbacks. Next, a modified version of each measure for reducing their estimation errors is introduced.

### 2.1 Distance Measures

For gap estimation, the BB method simply computes the length of a horizontal line between the smallest rectangles respectively enclosing two adjacent connected components, as shown in Fig. 1 (a). If the bounding boxes overlap horizontally, i.e. the right boundary of the left box extends to the left boundary of the right box, the distance is considered as zero. The RLEH method, shown in Fig. 1 (b), uses the minimum run-length or the minimum Euclidean distance with some heuristics to estimate the gap between a given pair of connected components. If the connected components overlap vertically by more than a threshold determined heuristically, the minimum run-length is used, and the minimum Euclidean distance, otherwise. In the CH method, the smallest convex polygon, called convex hull, enclosing each connected component should be estimated first. Next, we obtain the particular points where two adjacent convex hulls are intersected by the line linking their centroids. Finally, the gap between two connected components is defined as the Euclidean distance between these intersection points as shown in Fig. 1 (c).

These distance measures provide a simple gap estimation but frequently they produce serious estimation errors like underestimation or overestimation. The BB method suffers from underestimation due to the rectangular nature suppressing component shape to a box. As can be seen from Fig. 1 (a), the gap between the connected components, at least one of which has horizontally protruded ‘head’ or ‘tail’ part (see gaps **a**, **b**, **c**), is usually underestimated. The RLEH method estimates gaps directly from

component contours. In such case, estimation results are very sensitive to component shape and vertical positioning of components, as shown in Fig. 1 (b) (see gaps **a**, **b**, **c**). The CH method also suffers from an overestimation problem when an ascender or a descender is included in and located on either the beginning or the end part of connected component, as shown in Fig. 1 (c) (see gaps **a**, **b**, **c**).

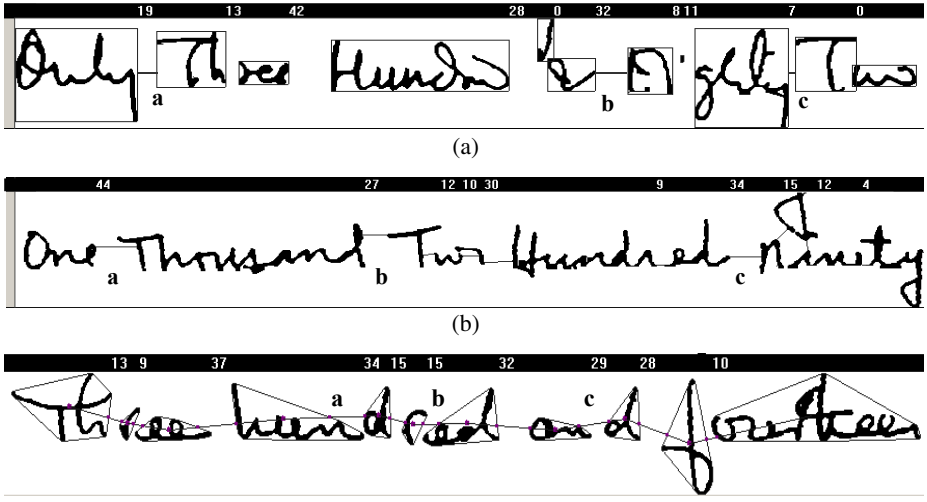



**Fig. 1.** Gap estimation using (a) BB, (b) RLEH, and (c) CH method

### 2.2 Modified Distance Measures

Next, we modify the above distance measures to avoid their overestimation or underestimation errors. In the case of BB method, we adjust the left and right boundaries of the bounding box of the component containing horizontally protruded 'head' or 'tail' part. Considering that such protruded 'head' and 'tail' parts usually consist of a single horizontal stroke as shown Fig. 2 (a), the 'head' part is defined as the region from the leftmost position to a node point (denoted by 'N') where the value of horizontal histogram is larger than $\beta W$. Here, $W$ is the average stroke thickness in the entire image calculated using a simple mathematical method proposed in [5], and $\beta$ is empirically determined as 1.25, considering some possible noise or distortion on a stroke. Similarly, the 'tail' part is defined as the region from the rightmost position to the node point (denoted by 'M'). Then, the new left and right sides of the bounding box are defined as follows:

$$L_x^{\text{new}} = L_x^{\text{old}} + \alpha H_{\text{wd}} \tag{1}$$

$$R_x^{\text{new}} = R_x^{\text{old}} - \alpha T_{\text{wd}} \tag{2}$$

Here, $H_{wd}$ and $T_{wd}$ denote the widths of the 'head' and 'tail' parts, respectively. And $\alpha$ is 0.35 if these parts are located within the body area and 0.25, otherwise. As can be seen from Fig. 2 (a), the gap between two connected components each of which has horizontally protruded 'head' or 'tail' part is quite well estimated by employing the modified bounding box (MBB).

The RLEH method is also modified (MRLEH) slightly based on such concept of adjusting bounding box to reduce its over-sensitivity to 'head' and 'tail' parts; the component contour located within adjusted bounding box is only considered for gap estimation. Unlike the MBB method, the parameter $\alpha$ is assigned as 0.25 or 0.2 according to the vertical positions of the 'head' and 'tail' parts. Furthermore, one more heuristic is introduced to deal with some special components like the dot in character 'i' or 'j', and broken part of a character; if two connected components do not overlap vertically, the gap between them is estimated by the bounding box distance instead of the run-length or Euclidean distance.

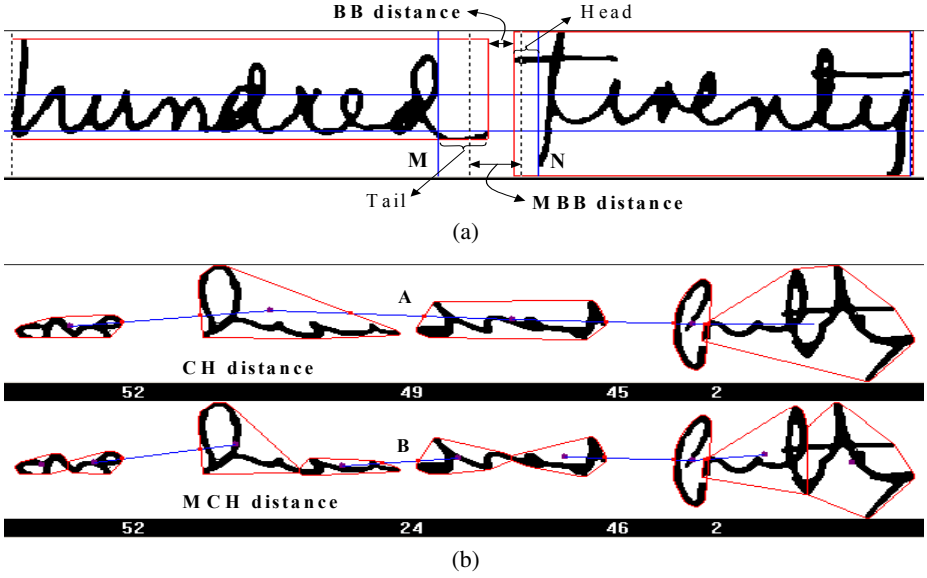



**Fig. 2.** Gap estimation using (a) modified BB distance and (b) modified CH distance

Lastly, in order to deal with the overestimation problem in the CH method, a pair of convex hulls for each connected component is introduced. We divide vertically a connected component into two equal parts and estimate two convex hulls enclosing each divided part. The gap between two adjacent components is then computed by considering two convex hulls enclosing the right part of the left component and the left part of the right component, as can be seen from Fig. 2 (b). This modified CH (MCH) method is quite attractive because it produces a reasonable gap distance even though an ascender or a descender is located on either the beginning or the end part of component whereby the gap is usually overestimated in the original CH method (compare gap **A** with **B**). Additionally, like the MRLEH method, a heuristic condition

for dot or broken part is also added to the MCH method; if two connected components do not overlap vertically, the gap is defined as the horizontal distance between the intersection points of two convex hulls, instead of the Euclidean distance.

## 3 Experimental Results

In experiments, we investigated the effectiveness of the modified distance measures by applying them to the word separation task using 1030 image samples of legal amounts included in CENPARMI database called IRIS and comparing their performance to that of the original measures. The IRIS database was obtained from real-life bank cheques. Thus considerable noise, shape distortion, and space irregularity were present in the images. Before performing word separation, several preprocessing procedures including smoothing, slant detection and correction, and removal of non-character components such as line, comma, and numeral parts were conducted in advance.

### 3.1 Word Separation by 2-Class Clustering

To extract words from a legal amount image based on the spatial gap distance, we employ a clustering technique based on the LBG algorithm [6] that classifies all gaps within a given legal amount image into two classes: inter-character gap (ICG) and inter-word gap (IWG). Before clustering for word separation, some special image samples containing only ICGs or only IWGs are extracted first according to the following heuristic procedure.

(1) For a given legal amount image, calculate three types (based on BB, RLEH, and CH methods) of the maximum, minimum, and average gap distances.
(2) If the width from the leftmost connected component to the rightmost one is less than 15% of the entire width of image, all gaps are assigned as ICG.
(3) If that width is less than 35% of the image width, and at least two types of the maximum gap distances are less than the overall gap average obtained from the whole data set or all three types of the average distances are less than 70% of the overall gap average, all gaps are assigned as ICG.
(4) If that width is larger than 35% of the image width and at least two types of minimum gap distances are larger than the overall gap average, all gaps are assigned as IWG.

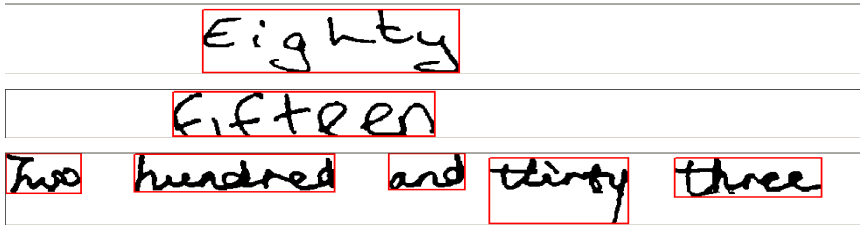
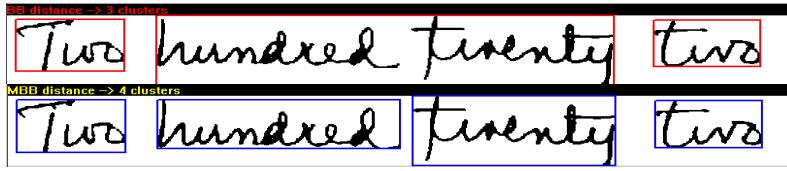



**Fig. 3.** Extracting words from image samples containing only inter-character gaps (ICG) or only inter-word gaps (IWG).

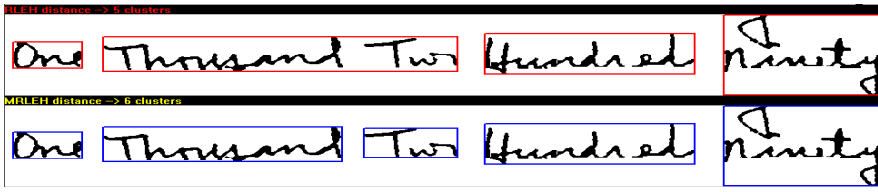
Figure 3 shows the examples of word separation for the image samples, in which only ICGs (1st and 2nd images) or only IWGs (3rd image) appear, using the above heuristic procedure. After extracting such special image samples, the clustering procedure based on LBG algorithm is then applied to the remaining samples. The experimental results in Table 1 show that the RLEH method performs best with its correct separation rate of 70.1% among the three original distance measures. Here, it should be noted that correct separation means that all words in a given legal amount image are perfectly isolated through the clustering procedure. Also, it can be found that all modified distance measures produced about 3% of better separation rate, when compared to their corresponding original distance measures.

**Table 1.** Experimental results of word separation

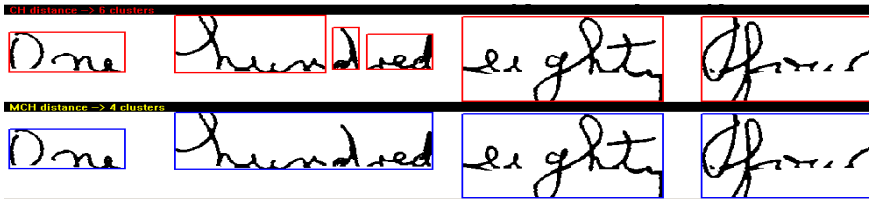
| Distance Measures | BB | RLEH | CH | MBB | MRLEH | MCH |
|---|---|---|---|---|---|---|
| Correct Separation (%) | 69.0 | 70.1 | 68.3 | 71.8 | 72.7 | 71.7 |




(a)




(b)




(c)

**Fig. 4.** Examples showing word separation error made by the original distance measures (upper part of each figure) and correction by their modified versions: (a) BB and MBB, (b) RLEH and MRLEH, and (c) CH and MCH

Figure 4 shows the examples of word separation illustrating the effectiveness of the modified distance measures. The words "hundred twenty" and "Thousand Two" shown in Fig. 4 (a) and (b) are misrecognized as one word (upper part of each figure) due to the underestimation by BB method and shape sensitivity by RLEH method, respectively. However, these parts are successfully divided into two words by using the modified measures, as shown in the lower part of each figure. In Fig. 4 (c), the word "hundred" is incorrectly divided into three parts due to the overestimation by the original CH distance but successfully merged by the modified method.

Next, we integrate these three different types of distance measures, thereby compensating the separation errors in each measure to achieve further improvement in word separation. A detailed description for our approach is provided in the following subsection.

### 3.2 Combining Three Distance Measures Based on 4-Class Clustering

The Venn diagrams shown in Fig. 5 represent the distribution of word separation errors due to the original BB, RLEH, and CH methods, and their modified versions, respectively. From these diagrams, it can be found that many errors are commonly produced from two or all of three distance measures even modified methods are applied. Thus a simple combining scheme such as the conventional majority voting is not expected to be effective in reducing such shared errors.

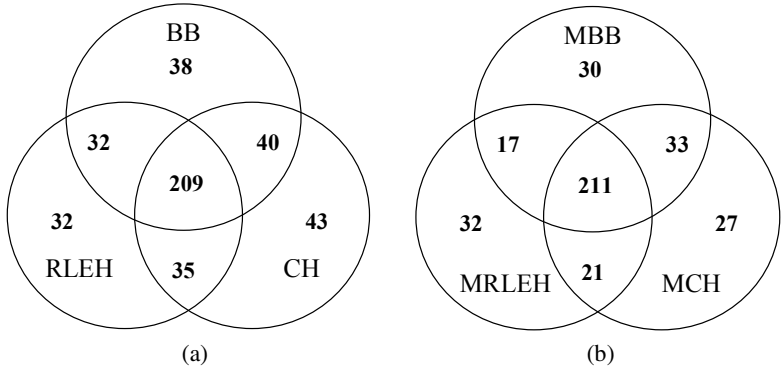



**Fig. 5.** Distribution of word separation errors in (a) BB, RLEH, CH methods and (b) their modified versions

In order to compensate effectively those errors commonly produced from two or all of distance measures as well as the errors caused by one measure only, we propose a new method of integrating three individual measures based on a 4-class clustering technique as follows:

(1) Estimate all gaps in a given image using a distance measure and divide them into 4 classes using LBG algorithm. The partitions are sorted in order of magnitude of their centroids.

(2) Assign an integer value, $\alpha \in \{2, 1, -1, -2\}$ to all the gaps according to the class to which they belong.

$$\alpha_i = \begin{cases} 2 & for \ \ g_i \in class\ 1 \\ 1 & for \ \ g_i \in class\ 2 \\ -1 & for \ \ g_i \in class\ 3 \\ -2 & for \ \ g_i \in class\ 4 \end{cases} \tag{3}$$

where, $g_i$ denotes *i*-th gap in a given legal amount image. Thus, a larger positive value ($\alpha = 2$) is assigned to the gaps belonging to the class with a smaller centroid (*class* 1) to accentuate their possibility of ICG, and vice versa.

(3) For the *i*-th gap, three different integer values, $\alpha_i^{BB}$, $\alpha_i^{RLE}$, and $\alpha_i^{CH}$ are assigned according to the distance measures: MBB, MRLEH, and MCH.

(4) Define *i*-th gap as ICG if $\alpha_i^{BB} + \alpha_i^{RLE} + \alpha_i^{CH}$ is positive, and IWG, otherwise.

In the second part of word separation experiments on the proposed combining method, we achieved more than 75% of correct separation rate. This performance is much higher than those of any types of individual distance measures so far used in our experiment as well as those of other studies that used a similar database for evaluating their approaches, as shown in Fig. 6.

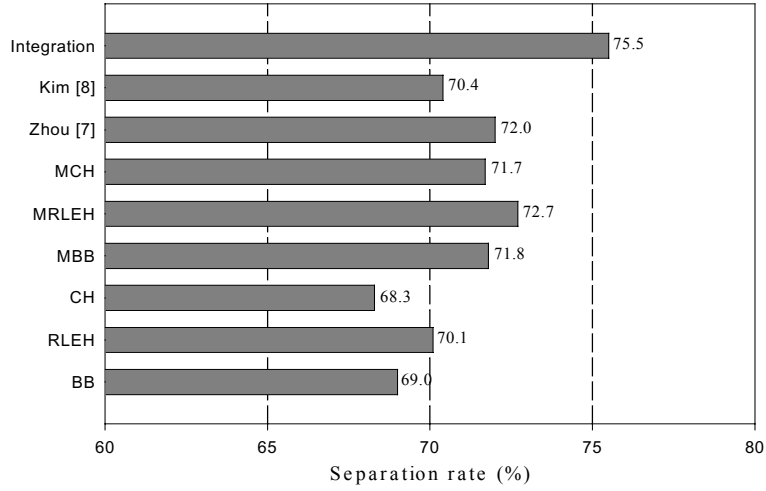



**Fig. 6.** Comparison of word separation rate according to methodologies

The Venn diagram in Fig. 7 clearly shows that the errors generated from only one of three distance measures are almost perfectly removed. Moreover, it can be found that the errors common to two of three distance measures are also reduced effectively. However, the number of errors common to all of three distance measures is hardly reduced even when the combining method based on the 4-class clustering is applied.

A primary assumption to use the spatial gaps between connected components for the problem of dividing a sentence into words is that the gaps between words are usu-

ally larger than those between characters. However, such assumption does not match well with the problem of extracting words from a legal amount on a bank cheque; an inter-word gap is not always larger than an inter-character gap, and horizontal overlapping and touching of two adjacent words arise frequently due to the writer's unconstrained writing style and writing space constraints. Accordingly, our analyses indicate that employing a prior knowledge such as possible maximum and minimum lengths of a word in lexicon or the number of possible words in a legal amount is needed to remove such troublesome errors. Moreover, introducing other methodologies such as implicit segmentation scheme or recognition based segmentation approach will be helpful to achieve further improvement in word separation performance.

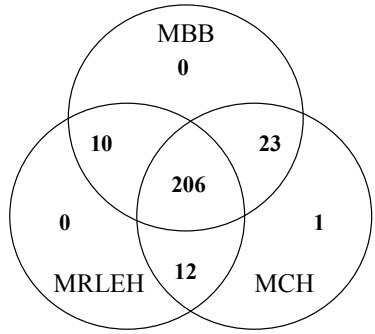



**Fig. 7.** Error distribution by combining method using 4-class clustering

## 4 Conclusions

Extracting words from the legal amount on a bank cheque has been performed based on the spatial gaps between connected components. The existing distance measures: BB, RLEH, and CH, are simple and quite efficient for gap estimation but all suffer from the inherent problem of underestimation or overestimation. To alleviate such problem, we have modified each distance measure; adjusted the left and right boundaries of the bounding box for the BB and RLEH methods, and introduced a pair of convex hulls enclosing equally divided connected components for the CH method. Furthermore, we proposed a salient method of integrating three individual measures based on 4-class clustering technique in order to effectively reduce the errors common to two of three distance measures as well as the errors made by each individual distance measure only.

Through a series of word separation experiments on CENPARMI IRIS database, we found that the modified measures show a better performance in terms of their separation rates compared with their corresponding original distance measures. Also, further improvement in performance of word separation was achieved by applying the combining method.

As a future study, we plan to introduce a priori knowledge about the lexicon of legal amount and to employ other methodologies such as implicit segmentation scheme or recognition based segmentation approach to reduce word separation errors caused by gap irregularity and overlapping or touching between words.

## References

1. Guillevic, D., Suen, C.Y.: Recognition of Legal Amounts on Bank Cheques. Pattern Analysis and Applications, 1(1) (1998) 28-41
2. Kaufmann, G., Bunke, H.: Automated Reading of Cheque Amounts. Pattern Analysis and Applications, 3(2) (2000) 132-141
3. Seni, G., Cohen, E.: External Word Segmentation of Off-line Handwritten Text Lines. Pattern Recognition, 27(1) (1994) 41-52
4. Mahadevan, U., Nagabushnam, R.C.: Gap Metrics for Word Separation in Handwritten Lines. Proc. Int'l Conf. Document Analysis and Recognition, 1 (1995) 124-127
5. Schürmann, J.: Document Analysis – from Pixels to Contents. Proc. IEEE, 80(7) (1992) 1101-1119
6. Linde, Y., Buzo, A., Gray, R.M.: An Algorithm for Vector Quantizer Design. IEEE Trans. Communications, COM-28(1) (1980) 84-95
7. Zhou, J., Suen, C.Y., Liu, K.: A Feedback-based Approach for Segmenting Handwritten Legal Amounts on Bank Cheques. Proc. Int'l Conf. Document Analysis and Recognition, (2001) 887-891
8. Kim, K.K., Kim, J.H., Chung, Y.K., Suen, C.Y.: Legal Amount Recognition Based on the Segmentation Hypotheses for Bank Check Processing. Proc. Int'l Conf. Document Analysis and Recognition, (2001) 964-967