http://dx.doi.org/10.5392/IJoC.2012.8.1.039

Character Segmentation in Chinese Handwritten Text Based on Gap and Character Construction Estimation

ChengDong Zhang, GueeSang Lee*
Department of Electronics and Computer Engineering
Chonnam National University, Gwangju, 500-757, Korea

ABSTRACT

Character segmentation is a preprocessing step in many offline handwriting recognition systems. In this paper, Chinese characters are categorized into seven different structures. In each structure, the character size with the range of variations is estimated considering typical handwritten samples. The component removal and merge criteria are presented to remove punctuation symbols or to merge small components which are part of a character. Finally, the criteria for segmenting the adjacent characters concerning each other or overlapped are proposed.

Keywords: *Character Segmentation, Projection Histogram, Handwritten Chinese, Character Construction Classification, Stroke Trace Estimation.*

1. INTRODUCTION

Offline recognition of connected handwritten characters plays an important role in many applications such as letter address reading, bank check reading, and fax document reading. For these kinds of recognition applications, character segmentation is an effective preprocessing system to increase the accuracy of character recognition. Incorrect segmentation will lead to false recognition[1], [2].

A major challenge for character segmentation in handwritten Chinese is that each Chinese character is usually a combination of pairs of isolated components[3]-[12]. These components may be written touching or overlapping with each other. However, usually each of these isolated components can also be an independent character.

There are seven different character combination styles (or structures), such as

'的' (symmetrical structure),
'住' (small-left and big-right structure),
'刮' (big-left and small-right structure),
'例' (left-middle-right structure),
'了' (independent structure),
'管' (upper-lower structure), and
'困' (center-focus structure).

Because of the different construction of various Chinese characters, their degree of written complexity is also different. As a result, various handwriting styles will be encountered. Some of the styles of writing '的' are shown in Fig. 1.

a) b) c) d) e)

Fig. 1. a) the standard writing style of '的', b)– e) different handwritten styles of '的'.

This Chinese character '的' is made by combining two isolated components, '白' (white) and '勺' (spoon). The handwritten styles shown in Figs. 1(b) and 1(c) are close to Fig. 1(a). In Fig. 1(d), there is a small and non-obvious gap between '白' and '勺'. However, in Fig. 1(e), the gap between the two components is wide and striking; this kind of situation will possibly lead to over-segmentation. Similar situations were encountered in other character constructions.

One of the requirements of Chinese writing is to write from left to right; this issue will be a major challenge for character segmentation. In addition, different levels of character complexity related to various complex situations can be found in a single sentence.

a)

b)

* *Corresponding author. E-mail : gslee@chonnam.ac.kr*
Manuscript received Aug.29, 2011 ; accepted Jan.09, 2012

c)

Fig. 2. Three Chinese handwriting styles.

Fig. 2 illustrates three handwriting styles for Chinese characters. In Fig. 2(a), each character is isolated. Fig. 2(b) shows two characters adjacent to each other, but without connection. Fig. 2(c) illustrates the situation where two pairs of characters are adjacent and connected to each other. Moreover, the size and construction of the adjacent characters are different.

Chinese character handwritings can result in different cases:

Each character is separated and isolated in the simplest case.

Some characters are composed of two or more separate components and these components are written apart.

Adjacent characters are written in such a way that bounding boxes of the characters overlap with each other or the characters may touch each other.

Because of these complex situations involving both single characters and pairs of characters, handwritten character segmentation is more complex than for printed characters.

Some previous research has been performed in this area using various methods. Chiang and Yu[13] and Kuo and Wang[14] proposed the method for the segmentation of handwritten Chinese characters, but when the neighboring characters touch each other, those methods fail to find the correct segmentation boundaries. Tseng and Chen [15] proposed a method in which strokes were extracted to build a stroke bounding box. Then knowledge-based merging operations were used to merge the stroke bounding boxes. Finally, a dynamic programming technique was used to find the globally optimal segmentation boundaries. Chen and Zhen [16] proposed a method based on hidden Markov models to produce segmentation paths and prune redundant paths. Finally, based on the gap ratio and longer-edge criterion, the optimal clustering scheme was determined. The experimental results proposed in [15] and [16] look good, but there is one issue not covered -- the punctuation. The punctuation usually is inserted between characters, so it can affect the accuracy of character segmentation and we need to consider this issue.

In this paper, a proposed method based on the analysis of variations in handwriting and character construction classification is used to segment offline handwritten Chinese text. First, a vertical projection histogram was used to locate the gap between each character and to gather information about character construction for basic segmentation. In the second stage, based on available information, a handwriting size range was estimated for different character constructions. Then the small components which make up a whole character were merged and remove the punctuation region. The merge criteria were based on the different character construction sizes estimated earlier. Finally, a segmentation method was applied to the miss-segmented area based on the estimated character construction size and on changes in the stroke trace.

2. SYSTEM OVERVIEW

A widely used technique for gathering information on characters is projection profile analysis. Based on available information, a first segmentation pass, called "basic segmentation," was performed, and two distance datasets representing gap space and possible character regions were constructed. These two distance datasets were used as segmentation criteria to support the last two stages of segmentation, the second stage for reducing over-segmentation and the last stage for obtaining the results.

The "basic segmentation" approach is discussed in Section 3.1. The gap and character distance datasets are also estimated in this phase. The criteria for removal of small components based on the two distance datasets are discussed in Section 3.2. In Section 3.3, merge criteria are discussed for certain miss-segmentation situations in which two components are each big enough to be a character, but should be merged into a single character. In Section 3.4, a merge method based on three kinds of criteria is discussed for adjacent characters within a width region. The last section presents experimental results and conclusions.

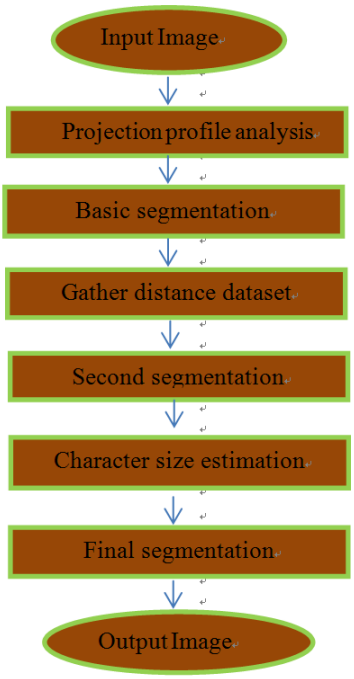


Fig. 3. Overall system flowchart for character segmentation.

3. PROPOSED METHOD

3.1 Projected profile histogram and basic segmentation

The projection histogram technique is used on a binary text image to gather pixel information based on projections in the horizontal and vertical directions. Here, only the vertical projection will be used. A projection on the vertical axis is obtained by adding up all the pixels row-wise:

$$P_{hor}(i) = \sum_{i} g(i, j)$$

Where, i is the column index of the input image, j is the row index of the input image, g(i,j) is the pixel value in the i-th column and j-th row, $P_{hor}(i)$ is the sum of the pixel values in the i-th column.

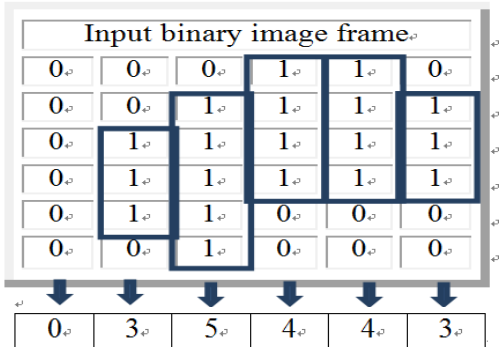
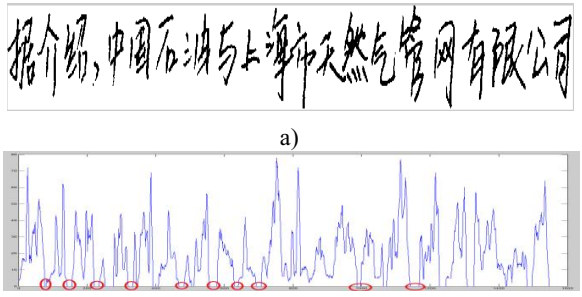


Fig. 4. Row sums used to obtain vertical projections.

Because they are derived from the HIT-MW dataset, all the input images are already binary images. To facilitate information gathering, the colors of the background and the foreground were first inverted. In addition, the vertical binary value was projected onto the vertical axis. The projection profile is shown in Fig. 5(b).



a)

b)

Fig. 5. a) Input image of handwritten Chinese, b) histogram of vertical projected profile.

Fig. 5(b) shows that in several locations, the value of the projection is zero. This means that the zero location represents a gap between two components. Based on the projection histogram, "basic segmentation" is performed by drawing vertical lines to divide the components at points where the projection value is equal to zero; the results of this basic segmentation are shown in Fig. 6.

Fig. 6. Basic segmentation of input image based on projected

After basic segmentation, two types of position nodes were located. The first type of node is located at the first point where the projection value is zero for each gap space and is called the "start node". The second type of node is located at the last point where the projection value is zero for each gap space and is called the 'end node.' Then, two types of dataset were constructed based on the 'start node' and the 'end node.' One consists of all the horizontal distances between each segmentation line and is called the 'possible character size range'. The other consists of all the horizontal distances for each segmentation line and is called the 'possible handwritten gap space.'

The 'possible handwritten gap space' (gap(i)) and the 'possible character size range' (wr(i)) can be calculated as follows, where pp is the projection profile histogram.

For i = 1 to length(pp)
{
If ($pp(i)$!= 0 && $pp(i+1)$ = 0) $node\_left = i$
If ($pp(i)$ = 0 && $pp(i+1)$!= 0) $node\_right = i$
$gap(i) = node\_left - node\_right$
$wr(i) = node\_right - node\_left$
}

3.2 Removal criterion based on distance dataset

The major goal of the second segmentation pass is to estimate variations in handwriting such as character size and character spacing. For the analysis of Chinese character structures, we consider the size of standard printed document first. Each character is composed of independent components. Some of Chinese characters can be used as the 'Chinese radical' which becomes a component of a character. At the same time, those characters can function as a 'Meaning unit' which can either be a single character or a main component of a character. The 'Chinese radical' is usually smaller than the 'Meaning unit' in size, when it is defined by the corresponding bounding box.

There are several '*Chinese radical*' samples with the corresponding '*Meaning unit*' as shown in Fig. 7.

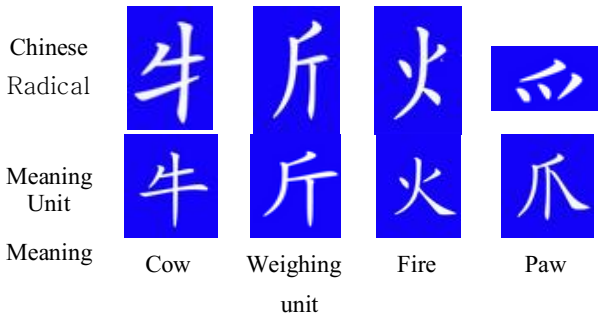


Fig.7. 'Chinese radical' and its corresponding 'Meaning unit'

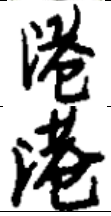
In a printed document text, the size of a character can be replaced by its horizontal length. In different structures of Chinese characters, the lengths of 'Chinese radical' and 'Meaning unit' appear differently as shown in Table 1, where 'Unit_R' indicates 'Chinese radical' and 'Unit_M' indicates 'Meaning unit'.

Table 1. Character sizes in printed document

Structure types	Size of a Character		
	Unit_R	Unit_M	The whole size
Symmetrical	0.0	1.0*2	2.0
Small-left	0.5	1.0	1.5
Big-left	0.5	1.0	1.5
Upper-lower	0.0	1.0	1.0
Left-middle-right	0.5*2	1.0	2.0
Centre-focus	0.5	1	1.5
Independent	0.0	1.0	1.0

Next, Table 2 shows some samples to compare with printed characters and handwritten characters.

Table 2. Different structures result in different sizes.

	Independent Structure	Small-left Structure
Printed character	了	港
Handwritten Samples		

According to samples shown in Table 2, it is easy to figure that the size of a handwritten character is sensitive to structures as well as to the complexity of a character. It means that in handwritten cases, "Unit_R" and "Unit_M" might be quite different from printed characters. Therefore, the size a handwritten character can better be represented by an interval considering the variations.

Based on the analysis of variations in handwriting, the size of a handwritten character can be represented by a linear system:

$$H_{size} = a * D_{size} + n$$

where, H_{size} is the expected size of the handwritten Character;
D_{size} is the size of the printed character;
a reflects the difference between printed characters or components and their handwritten ones,
n is intra-gap parameter.

According to the linear system, the intra-gap variation is represented by a number n, defined less than 0.5. If the *mean value* of 'possible *character size range'* set as the $a * D_{size}$, where the parameter a is an interval between the smallest size and the biggest size. The biggest size of a printed character is double compared to the smallest character, considering noise.

So H_{size} can be represented as an interval $[s*m, b*m]$, where s and b indicate the smallest size and the biggest size respectively and m represents the mean value of characters. Now the parameters of s and b can be decided with some heuristic, which are 0.7=(standard size/2 + noise/2) to 1.5=(double size + noise), assuming the standard size is one unit length and noise can be one half of the unit, in our approach. It means that if the size is bigger than *mean value* * 1.5, there are at least two characters contained in this region. Also, if the size is smaller than mean value * 0.7, we need to estimate whether this region is a punctuation region or not.

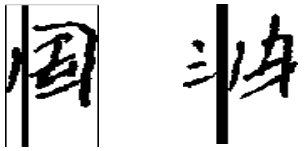
A pre-processing step consists of calculating the mean values of '*possible handwritten gap space*' and '*possible character size range*'.

Sum the regions which size as follows:

$$wr(i) > \min(wr) \;\&\&\; wr(i) <= \text{mean}(wr) * 1.5$$

And, calculate the mean value of these regions as the new threshold T.

The new threshold means that this is a weighting value for a middle size possible character region, excluding punctuation regions and unsegmented regions.



a) b)

Fig. 8. a) Result of over-segmentation of '国', (b) result of oversegmentation of '油'.

The challenge for the removal of punctuation is that the size of a simple character such as '了' and punctuations are similar, for which the ratio of character pixels and all the pixels in the region are considered. The punctuation just occupies a small space in the bounding box or the limited region. We computed the ratio between character and limited region for simple characters in Chinese character such as '了' and '人', which shows that each character at least occupies 40% of its limited region while the punctuations occupy much less. Based on this observation, the removal criterion of punctuation region can be summarized as follows:

(1) Compute all $wr(i)$ which are smaller than $min(wr)/2$.
(2) If the component pixels occupy less than 40% of the limited region, remove wr (i).

Sample results are shown in Fig. 9.

a) b)

Fig. 9. The removal result of punctuation region

According to the analysis of handwritten character size in previous session, these constructions can be divided into two groups. The first group, called the '*small-size group*' contains the 'independent structure,' 'big-left and small-right structure,' 'upper-lower structure,' and 'small-left and big-right structure'

styles of structures. The second group, called the 'big-size group,' contains the 'centre focus structure,' 'symmetrical structure,' and 'left-middle-right structure' styles of structures. The characters shown in Fig. 8 clearly all belong to the 'small-size' group.

The new threshold is the middle value between the '*small-size group*' and the '*big-size group*'. Two specific thresholds were calculated using the new threshold as follows:

$$sg = T * 0.7;$$
$$bg = T * 1.5;$$

Where, *sg* is the range threshold of the '*small-size group*' and *bg* is the range threshold of the '*big-size group*'.

The two characters '国' and '油' have different constructions and gap spaces, but the size of the two characters indicates that they belong to the '*small-size group*'.

To resolve this kind of issue, a gap space threshold should also be estimated based on the '*possible handwritten gap space*' dataset. As shown in Fig. 8, some characters lie adjacent to each other without any gap space. Moreover, the gap space between '限' and '公' is evidently smaller than the other gaps. To deal with this kind of issue, the gap threshold *gt* of the '*possible handwritten gap space*' dataset was calculated excluding small gap spaces such as in Fig. 10.

Fig.10. Small gap space between '限' and '公'.

The next step is, based on the range thresholds *sg*, *bg* and the gap threshold *gt*, to remove certain components such as radicals from whole characters, as shown in Fig. 8. The small radical components will be removed, as will some kinds of over-segmentation. The proposed criterion is the following:

```
If (wr(i) < sg & gap(i) < gt & gap(i-1) < gap(i) & gap(i-1) < gt)
{
        Remove gap(i-1);
}

If (wr(i) < sg & gap(i) < gt & gap(i-1) > gap(i) & gap(i) < gt)
{
        Remove gap(i);
}

If (wr(i) < sg & gap(i) < gt & gap(i-1) = gap(i) )
{
    If ((wr(i) + wr(i-1) < wr(i)+wr(i+1)) &
            (wr(i) + wr(i-1) < bg) & (wr(i) + wr(i+1) < bg))
        Remove gap(i-1);

    If (wr(i) + wr(i-1) > bg & wr(i) + wr(i+1) < bg)
        Remove gap(i);

    If (wr(i) + wr(i-1) < bg & wr(i) + wr(i+1) > bg)
        Remove gap(i-1);
}
```

Based on the removal criterion, over-segmented areas were removed as shown in Fig. 11.

Fig.11. Second phase of over-segmentation removal.

3.3 Merge criteria based on distance dataset

In the previous section, it was stated that the single character '的' can be written in various styles. The main issue arises for 'symmetrical constructions,' such as the '的' shown in Fig. 1. Because of the symmetrical nature of this character, the size of the two parts of '的' is almost the same, but the difference in gap space between the two parts is evident. A similar issue is illustrated in Fig. 12.

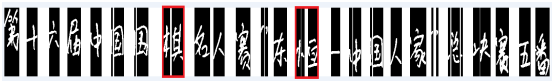


Fig.12. Over-segmentation of characters '棋' and '恒'.

The merge criterion for this issue, also based on the range thresholds *sg*, bg and the gap threshold *gt*, is used to merge two components which in fact belong to a single character. The fine-merging operation is performed as follows:

```
If ( (wr(i) < sg   + wr(i+1) <= bg *150%)   &   (gap(i) < gt) )
        { Merge wr(i) and wr(i+1); }
```

Based on the removal and merge criteria, the fine-merging result is as shown in Fig. 13.

Fig.13. Combined result based on the removal and merge criteria.

3.4 Adjacent-region segmentation based on the triple criterion

Because of individual variations in handwriting, a part of one character may be inserted into or adjacent to another character. As a result, serious miss-segmentations may occur.

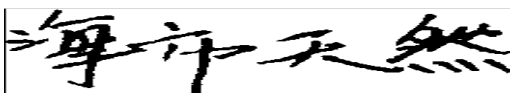


Fig.14. Miss-segmentation result of '海', '市', '天', and '然' because of adjacency or insertion of another character.

Usually, the handwriting style of a human being is very hard to change. In the step described in the previous section, based on the two distance datasets, two range thresholds for 'small-size group' and 'big-size group' were generated. These two range thresholds also play a major role in the triple criterion.

Another important rule is based on changes in the trace of handwritten strokes. Because changes of strokes in the vertical direction are not considered important, the position of each stroke from bottom to top in each column is determined. When the first black pixel position is located, the algorithm stops and jumps to the next column. The trace map corresponding to Fig. 11 is shown in Fig. 15.

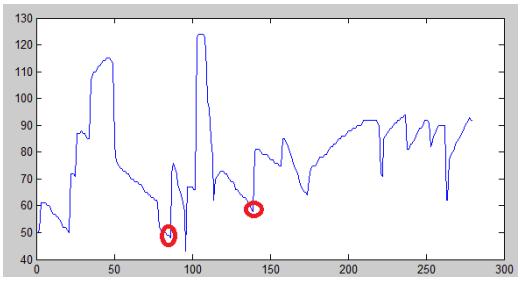


Fig.15. Stroke trace maps of '海', '市', '天', and '然'.

The size of this miss-segmentation region is 280. The peaks and gaps in the stroke map represent how the stroke is written and the transition to the next stroke; because the direction and position of the two strokes are different, the change is evident. The position of peaks and gaps is the second criterion for adjacent region segmentation.

The third criterion is called 'break-point matching.' Even though two characters may be adjacent to or inserted into each other, the strokes in the adjacent region are normally very weak in the vertical direction compared to the main part of the character. A search is therefore conducted for weak points in the projection histogram for this limited region. The reader is invited to compare Fig. 15 and Fig. 16.

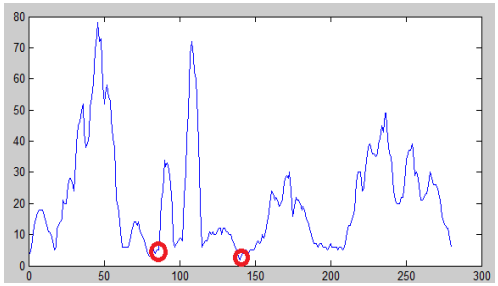


Fig.16. Projection map of limited regions of '海', '市', '天', and '然'.

It is apparent that the two positions marked in the two figures are matched in the same location. This means that these two positions exhibit evident stroke change and weak connection. This location is the desired segmentation point. Moreover, in the region from 180 to 280, a weak connection area was located in Fig. 16. Unfortunately, no corresponding peak or gap was located in Fig. 15. To resolve this problem, the third criterion was combined with the first criterion. Based on these two criteria, the possible character size was estimated first. Then several markers were added to the miss-segmentation region from 180 to 280, which helped to locate a final position based on the information in Fig. 16.

In the triple-criterion system, the ideal situation is that the point in question satisfies all three criteria; if not, at least two criteria should be satisfied. The final result is shown in Fig. 17.

Fig.17. Final segmentation result after miss-segmentation region removal.

4. EXPERIMENTAL RESULTS

For experiment, we have used 400 samples taken from HIT-MW dataset[17], [18]. The algorithm has been implemented in Matlab and run on Windows in Quadcore flatform. Fig. 18 shows some sample results and Table 3 shows the comparison of other method and our proposed method. In Fig. 18, the experimental results show that the proposed method works well for given handwritten sentences, even for the sentence with a skew angle direction. The removed results of punctuation region are marked by red rectangles. The major difficulty is that for certain small scattered and components such as '、' in '心' and '拨'. In some situation, the components '、' in handwriting will be located far from the main body of the character. Moreover, this component is very similar to a comma. In this case, the component '、' is easy to merge with other components. Even, it is very hard to recognize the different between this kind of component and punctuation. The location marked by blue rectangle means that the small component should belong to the character. But it has been removed as the punctuation region. Another weak point of the proposed method may not be effective for the numerical digits. We can see in the experimental results that the segmentation result for the numbers does not work well. The proposed method is based on the analysis of Chinese character structures and the variations in handwriting of Chinese people. Obviously, this analysis may not offer solution to the segmentation of numerical characters. The number region is marked by green rectangles to show that the segmentation of numbers fails sometimes.

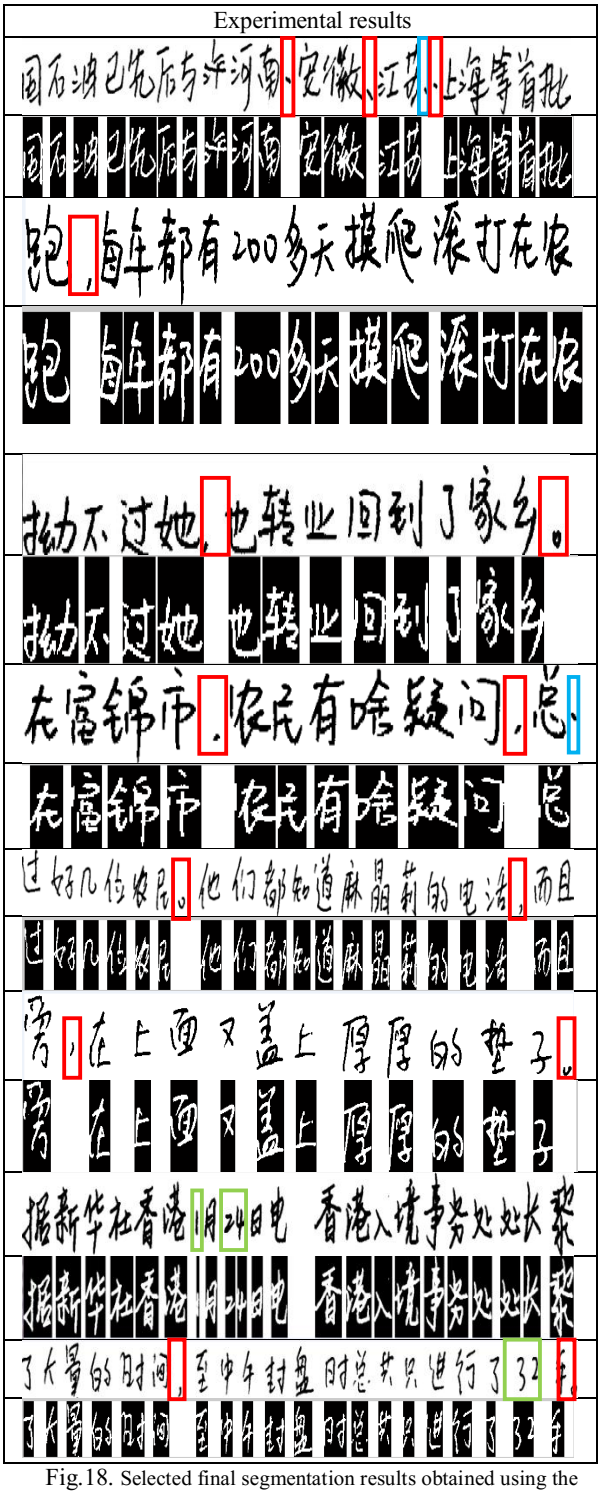


Fig.18. Selected final segmentation results obtained using the proposed method.

Table 3. Comparison with other method and our proposed method

	Character Count	Accuracy (%)
Method in [16]	649	88.75
Our Method	400	94.5

5. CONCLUSIONS

In this paper, a segmentation method for Chinese handwritten characters is presented. The method is based on a projection profile histogram and uses two distance datasets to estimate character construction for segmenting most handwritten Chinese characters effectively. Dealing with small scattered components and numbers will be the focus of future works.

ACKNOWLEDGEMENT

This research was supported by Basic Science Research Program through the National Research of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0006109) and the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2012-H0301-12-3005).

REFERENCES

[1] Y. Lu. "Machine printed character segmentation: an overview," *PR journal, vol.28,* 1995, pp. 67-80.

[2] Y. Lu and M. Shridhar. "Character segmentation in handwritten characters: an overview," *PR journal, vol.29,* Sep, 1996, pp. 77-96.

[3] L.Y. Tseng, R. C. Chen. "Segmenting handwritten Chinese characters based on heuristic merging of stroke bounding boxes and dynamic programming," *PRL journal, vol.19,* Oct, 1998, pp. 963-973.

[4] Q.S Chen, L.X Zhen. "Character segmentation in handwritten Chinese text image based on component clustering techniques," *Proc . TENCON '02*, pp. 435-440.

[5] Y. Jiang, X. Ding, Z. Ren. "Substring Alignment Method for Lexicon-Based Handwritten Chinese String Recognition and its Application to Address Line Recognition," *Proc. ICPR'06*, 2006, pp. 683-686.

[6] C. Hong, G. Loudon, Y. Wu, et al. "Segmentation and Recognition of Continuous Handwritten Chinese Text," *PRAI journal, vol.12*, Feb.1998, pp. 223–232.

[7] Y.H Tseng, H.J Lee. "Recognition-Based Handwritten Chinese Character Segmentation Using a Probabilistic Viterbi Algorithm," *PRL journal, vol.20*, Aug.1999, pp. 791–806.

[8] J. Gao, X. Ding, Y. Wu. "A Segmentation Algorithm for Handwritten Chinese Character Strings," *Proc. ICDAR'09*, 1999, pp.633–636.

[9] S.Y Zhao, Z.R Chi, P.F Shi, et al. "Two-stage Segmentation of Unconstrained Handwritten Chinese Characters," *PR journal vol.36*, Jan, 2003, pp.145–156.

[10] N. Ezaki, M. Bulacu, and L. Schomaker, "Text detection from natural scene images: towards a system for visually impaired persons," *Proc. ICPR'04*, 2004, pp.683-686

[11] R.G Lasey and E. Lecolinet, "A survey of methods and strategies in character segmentation," *Tran. PAMI'06, vol.18, no.7,* 1996, pp.690-706.
[12] G. Seni and E. Cohen, "External word segmentation of off-line handwritten text lines," *PR journal, vol.27*, 1994, pp. 41-52.
[13] C.C Chiang and S.S Yu, "An iterative character segmentation method for irregularly formatted Chinese documents," *Proc. OCRDA'96*, 1996, pp.61-67.
[14] H.H Kuo and J.F Wang, "A new method for the segmentation of mixed hand printed Chinese/English characters," *Proc.ICDAR'93*, 1993, pp. 810-813.
[15] L.Y. Tseng, R. C. Chen. "Segmenting handwritten Chinese characters based on heuristic merging of stroke bounding boxes and dynamic programming," *PRL journal, vol.19,* Oct, 1998, pp. 963-973.
[16] Q.S Chen, L.X Zhen. "Character segmentation in handwritten Chinese text image based on component clustering techniques," *Proc . TENCON '02*, pp. 435-440.
[17] T.H Su, T.W Zhang, et al. "Corpus-based HIT-MW database for offline recognition of general-purpose Chinese handwritten text," *DAR journal, vol.10, no.1*, Oct, 2007, pp. 27-38.
[18] T.H Su, T.W Zhang and D.J Guan. "HIT-MW dataset for Offline Chinese Handwritten Text Recognition," *Proc. ICFHR'06*, 2006.

ChengDong Zhang

He is currently a MS student in Computer Science department of Chonnam National University, Korea. His interesting researches are Image processing, especially document analysis, text detection, text localization, text binarization and computer vision.

GueeSang Lee

He received the B.S. degree in Electrical Engineering and the M.S. degree in Computer Engineering from Seoul National University, Korea in 1980 and 1982, respectively. He received the Ph.D. degree in Computer Science from Pennsylvania State University in 1991. He is currently a professor of the Department of Electronics and Computer Engineering in Chonnam National University, Korea. His research interests are mainly in the field of image processing, computer vision and video technology.