Segmentation of Handwritten Characters for Digitalizing Korean Historical Documents

Min Soo Kim[1], Kyu Tae Cho[1], Hee Kue Kwag[2], and Jin Hyung Kim[1]

[1] CS Div., Korea Advanced Institute of Science and Technology,
373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701, Republic of Korea
{mskim,ktcho,jkim}@ai.kaist.ac.kr
[2] Dongbang SnC Co., Ltd.
10th Floor, BaekSang Bldg., Gwanhun-dong, Jongno-gu, Seoul, Korea
hkkwag@dbmedia.co.kr

Abstract. The historical documents are valuable cultural heritages and sources for the study of history, social aspect and life at that time. The digitalization of historical documents aims to provide instant access to the archives for the researchers and the public, who had been endowed with limited chance due to maintenance reasons. However, most of these documents are not only written by hand in ancient Chinese characters, but also have complex page layouts. As a result, it is not easy to utilize conventional OCR(optical character recognition) system about historical documents even if OCR has received the most attention for several years as a key module in digitalization. We have been developing OCR-based digitalization system of historical documents for years. In this paper, we propose dedicated segmentation and rejection methods for OCR of Korean historical documents. Proposed recognition-based segmentation method uses geometric feature and context information with Viterbi algorithm. Rejection method uses Mahalanobis distance and posterior probability for solving out-of-class problem, especially. Some promising experimental results are reported.

1 Introduction

Korea national agencies have been preserving the archives of Lee Dynasty contains over 2 million books with more than 400 million pages in printed format. However, it is common sense that the best way to preserve valuable documents is to reduce the frequency of their being physically accessed. As a result, the maintenance causes limited access to those documents. By the way, lots of people believe the construction of digital library for historical documents can expand accessibility.

As the digitalization is expected to enlarge utilization historical documents efficiently, several institutions have been digitalizing historical documents from several years ago. For instance, we were experienced in the digitalization effort took 7 years by manual key-in about Korean Canon of Buddhist which is made up of wooden printing blocks carved by hand, consists of 160,000 pages and

S. Marinai and A. Dengel (Eds.): DAS 2004, LNCS 3163, pp. 114–124, 2004.

Fig. 1. An Example of Korean Canon of Buddhist.

56,000,000 characters. The following Fig.1 shows an example of Korean Canon of Buddhist. However, in this case, the digital archives were constructed by complete manual approach.

Digitalization of historical documents have been mostly performed by manual typing, in Korea. Because most of historical documents were hand-written in ancient Chinese characters which we call Hanja. Ancient characters which are not used in comtemporary texts take considerable proportion, too.

(a) variants (b) Blurred characters

Fig. 2. Technical challenges for historical character images.

Recently, the use of OCR technique are getting attention because the latest OCR techniques show high performances on modern printed materials. But digitalization of Hanja documents is not so easy with just OCR technique. Its main difficulty comes from shape variation due to writers' habits, styles, and so on(Fig.2(a)). Also, blurred characters are often appeared in the documents because of the complex structure of its strokes and brush ink (Fig.2(b)). According to the facts deteriorating the performance of character recognition, it is impossible to expect the perfect output of OCR, and consequently, it cannot simply substitute for manual typing. Also the utilities of OCR to historical documents were very restricted[1][2]. As a result, we have developed a OCR-based system to enhance the overall efficiency of digitalization process[3](Fig.3). This system was designed for the combination of both typing and OCR to compensate one's drawback by others.

A brief description of our overall system is as follows: As the segmentation step is performed, scanned several hundred pages of documents are segmented into individual characters rapidly. Then OCR module is called up to classify each of the segmented characters. Namely, each segmented character image is fed to a recognizer to get a label. Segmentation and classification are repeatedly performed until all documents are processed. Also, remarkable characteristic of proposed system is to consider the rejection whenever character images are recognized. Namely, the system rejects if OCR confidence is not high enough. After

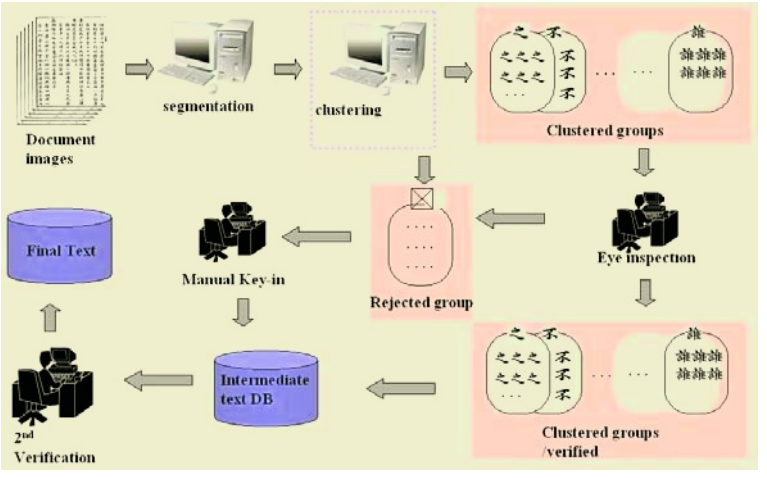

Fig. 3. Our overall digitalzing Scheme of Korean historical documents.

the classification and the rejection are done, the characters with the same class label are collected into one of predefined character groups. And the system shows them to operators to verify the result. Whenever the operators find misclassified character, they can remove that character. It can be found a set of less similar characters at the back from sorted grouping result of each character(Fig.4).

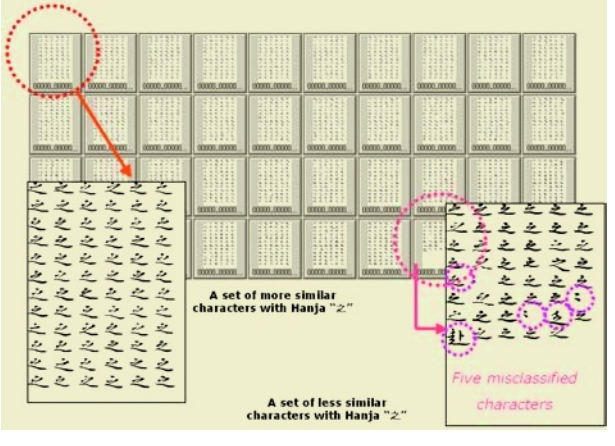

Fig. 4. User interface for eye inspection(Grouping about characters with same label).

Therefore, operators have only to investigate misclassified characters at the back part of grouped characters. When the verification is over, the label of each character in the group is automatically assigned by the system. By virtue of provided user interface for eye inspection, it saves lots a laborious typing effort.

This paper is organized into 4 sections. Section one is this Introduction in which we describe overall digitalizing system with the situation of Korean historical document digitalization. Section two and three explain about proposed segmentation method and rejection method which play a critical role in our OCR process, respectively. Section four shows the experimental results and evaluations. Finally, section five is the conclusion and prospectus of future work.

2 Proposed Recognition-Based Segmentation Method

Approach of Hanja segmentation can be classified into recognition-based method or image-based methods. Because image-based methods are highly dependent on character shape and handwritten characters have diverse variation, such techniques have a performance limitation for segmentation of handwritten characters. For automation of digitalizing historical documents, high accuracy of segmentation rate is essential. Therefore, image-based segmentation methods are not suitable for historical documents.

We propose to use recognition-based methods because they are effective in dealing with handwritten characters. But it is known that recognition-based Hanja segmentation methods have some problems [6][7]: (1) Out-of-class which has incomplete shape is matched into a character defined by recognition model. (2) It is time consuming to evaluate many cnadidates for merging character fragments. (3) Each fragment in a character can be misclassified as individual character (Fig.5).

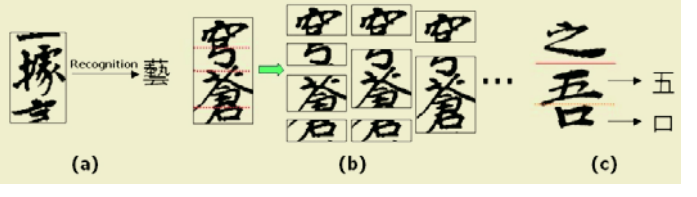

Fig. 5. Problems of recognition-based Hanja segmentation (a) Recognition of Out-of-class (b) Overhead due to recognition of many candidates (c) Misclassification of each fragment in a character into individual character.

In order to solve the problems associated with recognition-based segmentation methods, two additional criteria, character geometric feature and context information, are applied. Character geometric feature helps to reduce number of candidates for recognition and this reduction can decrease possibility of recognition for out-of-class and time needed for recognition. Context information can reduce the possibility of misclassifying a character fragment as individual character.

Because recognition-based methods employ a split-merge strategy in which the split segments are merged into a character, pre-segmentation stage to separate the text string image into segments is needed. Each segment must belong

Fig. 6. Pre-segmentation by nonlinear segmentation path.

to just one character. Nonlinear segmentation path is split method that can separate overlapping characters and touching characters [8].

For generating nonlinear segmentation paths, a character string is regarded as a multi-stage directed graph. Observation score for each pixel and transition score from left pixel to right pixel are defined in advance. The less number of black pixels a segmentation path passes and the more straight it is, the higher score it gets. Some possible paths from left end to right end of character string are selected by using Viterbi algorithm. As shown in Fig.6 nonlinear segmentation path can separate overlapping parts and touching parts. A segmentation graph is then constructed using candidate paths to represent nodes and combining scores to represent arcs. Posterior probability for geometric feature and character string giving character image is used for merging score. If the number of characters is k in a character string, combining candidates that are made up of splitted components are defined by $X = x_1, x_2, \ldots, x_k$, character string $S = s_1, s_2, \ldots, s_k$, and character geometric feature $G = g_1, g_2, \ldots, g_k$. Posterior probability $P(G, S|X)$ can be derived as follows:

$$\begin{aligned} S &= \arg\max_s P(G, S|X) = \arg\max_s P(G, S, X) \\ &\approx \arg\max_s P(G)^{\lambda_G} P(S, X) = \arg\max_s P(G)^{\lambda_G} P(S) P(X|S) \\ &= \arg\max_s \lambda_G \log P(G) + \log P(S) + \log P(X|S) \end{aligned}$$

In above equation, λ_G is a revision value because geometric feature G depends on character string and character image. This equation is derived as sum of character geometric feature P(G), context information score P(S) and recognition score $P(X|S)$. The constant λ is defined by training. Geometric score P(G) is defined as follows.

$$P(G) = P(CH)^{\lambda_c} P(SQU)^{\lambda_s} P(GAP)^{\lambda_p} \tag{1}$$

In (1), CH,SQU and gap denote the character height, squareness and internal gap respectively. Also, three constants $\lambda_C, \lambda_S, \lambda_G$ are assigned to 1,1,3, respectively. Distributions of CH, SQU and gap are estimated by Parzen window method. If P(G) is 0, the merging candidate will be eliminated. Context information score can be computed by bi-gram language model score between previous and current character labels in the character string. Recognition score is the distance between recognition model and merging candidate image. Finally, the segmentation graph is generated as shown in Fig.7. In Fig.7, merging candidate

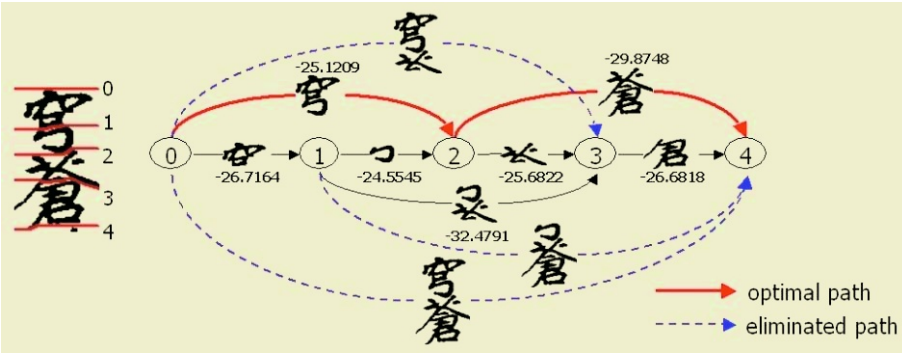

Fig. 7. Optimal segmentation path finding.

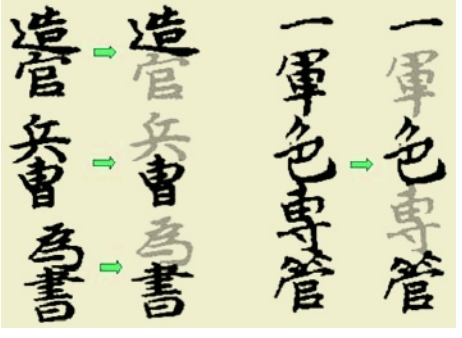

Fig. 8. Examples of segmented images by proposed method.

image without combining score is eliminated by character of geometric score. This elimination can reduce overhead of recognition. Some segmented images by searching optimal path are shown in Fig.8.

3 Recognition with Rejection-Option by Linear Discriminant Analysis

LDA(linear discriminant analysis) is a classification method using discriminant function based on Mahalanobis distance on the assumption that features are from normal random variables and covariances of each class are the same. Then, we can calculate conditional probability of $\mathbf{x}$ given w_i,

$$p(\mathbf{x}|w_i) = (2\pi)^{-n/2}|\Sigma|^{-\frac{1}{2}}\mathbf{exp}[-1/2(\mathbf{x}-\mu_i)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_i)]$$

We would allocate new observation to the population for which $P(w_i|x)$ is largest. Namely, we classify $\mathbf{x}$ to w_i if $P(w_i|\mathbf{x}) > P(w_j|\mathbf{x})$ for all $i \neq j$.

$$p(w_i|\mathbf{x}) = \frac{p(\mathbf{x}|w_i)p(w_j)}{P(\mathbf{x})}, p(\mathbf{x}) = \sum_{k=1}^{g} p(\mathbf{x}|w_k)p(w_k)$$

Because Bayes's classification rule depends on posterior probability $p(w_i|\mathbf{x})$, it is proportion to $\log p(\mathbf{x}|\pi_i) + \log p(\pi_i)$,

$$\begin{aligned} p(\pi_i|\mathbf{x}) &\propto \log p(\mathbf{x}|\pi_i) + \log p(\pi_i) \\ &= -\frac{1}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}(\mathbf{x}-\mu_i)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_i) + \log p(\pi_i) \end{aligned}$$

Since $\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}$ and $\log|\mathbf{\Sigma}|$ are common factors, this results in

$$p(\pi_i|\mathbf{x}) \propto 2{\mu_i}^T\Sigma^{-1}\mathbf{x} - {\mu_i}^T\Sigma^{-1}\mu_i + 2\log p(\pi_i)$$

If we substitute parameters to MLE(maximum likelihood estimator), we obtain LDF(linear discriminant function) like this.

$$d_i(\mathbf{x}) = 2\widehat{\mu}_i^T\widehat{\mathbf{\Sigma}}^{-1}\mathbf{x} - \widehat{\mu}_i^T\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mu}_i$$

In historical character classification with OCR, a very important problem is how many character classes are chosen as a relevant set in recognition. All Hanja classes don't need to be necessary in the recognition of documents, because many characters rarely appear in common use and this greatly increases the computational complexity of the recognition. We have statistically investigated the ancient Korean documents about 3,896 pages containing about 1.5 million characters over 5,599 Hanja classes of Seungjungwon Diary vol. 29. From the investigation, we observed that about 2,500 Hanja classes were frequently used, but about 3,600 Hanja classes were rarely used. Thus, we determined that 2,568 Hanja classes, which frequently appear about 99% in the documents, should be considered in the recognition step(refer to Fig.9). This is one reason we propose the rejection system. Another one is because the cost of detecting and correcting misclassified characters is more expensive than the cost of manual typing. In this paper, we propose the rejection method using maximum posterior probability threshold which removes ambiguous characters and using Mahalanobis distance threshold which throws out outliers together. Fig.11 shows results of rejection-used recognition. if we assume $\mathbf{x}|w_i$ represent a random samples from

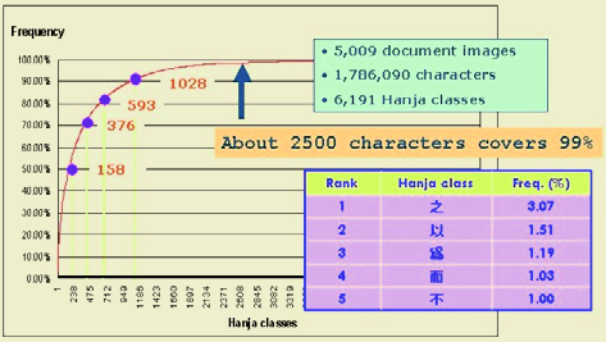

Fig. 9. The number of characters can be considered from frequency analysis.

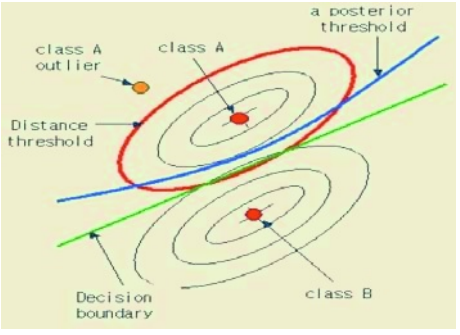

Fig. 10. Boundaries of two rejection rules.

Fig. 11. Rejection-used recognition results.

multivariate normal distribution with mean vector μ_i and covariance matrix $\mathbf{\Sigma}$. $P(w_i|\mathbf{x})$ can be calculated

$$P(w_i|\mathbf{x}) = \frac{exp(-\frac{1}{2} \times {r_j}^2)}{\sum_{k=1}^{g} exp(-\frac{1}{2} \times {r_k}^2)},\ r_j = ((\mathbf{x}-\mu_j)^T \mathbf{\Sigma}^{-1} (\mathbf{x}-\mu_j))^{\frac{1}{2}}$$

The system reject if maximum posterior probability is less than threshold θ_1 or shortest Mahalanobis distance is larger than threshold θ_2(refer to Fig.10).

Reject $\mathbf{x}$, if $\max\ p(w_j|\mathbf{x}) < \theta_1$ or $\min\ r_j > \theta_2$.
Accept $\mathbf{x}$, otherwise

4 Experimental Results and Evaluations

We evaluated the effectiveness of the proposed methods using Seungjungwon Diary, an ancient Korean government document, written by many writers during nearly 500 years. To show the performance of proposed segmentation method, were used 200 historical document pages that contain 78,756 handwritten characters of Seungjungwon Diary vol. 29. As may be shown in Fig.12, the performance of proposed segmentation method is nearly similar to that of manual segmentation. The recognition rates by manual and proposed segmentation methods were 92.99% and 92.98%, respecively. In comparison with manual segmentation result, proposed method achieved performance of 99.98%.

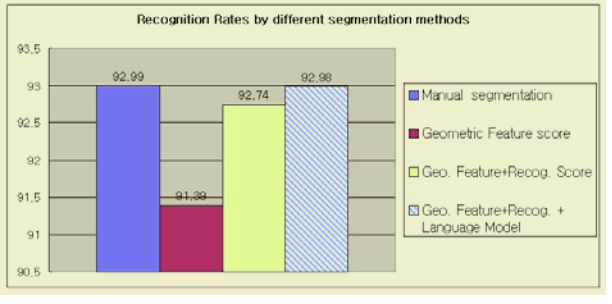

Fig. 12. Comparison of recognition between manual segmentation and proposed segmentation method.

Next, we carried out an experiment to compare precisions of the classifiers according rejection rates. 100 characters per class extracted from Seungjungwon Diary vol. 29 were used for training. Also, 200 pages extracted from Seungjungwon Diary vol. 29 were used for testing. E-classifier and M-classifier mean classifiers based on Euclidean distance and Mahalanobis distance, respectively. Also, d-thr and p-thr are Mahalanobis distance threshold and posterior probability threshold, respectively. As may be seen in Fig.13, On the whole, the M-classifier with d-thr and p-thr rejecter is superior to others.

As in Fig.13, when the precision(recognition rate) of accepted characters is 98%, we can see the percentage of rejected characters is 12.68%. From this result, table 1 shows the economical effectiveness of proposed system. We compared the input cost by manual typing with that of proposed system. Suppose that input cost and reform cost are 10 and 30, respectively, and we have 1,000,000 characters. When the ratio of rejected characters is 12.68%, total input cost of using proposed system is 18,680,000 by contrast with total cost of manual typing only method is 100,000,000.

Also, we calculated the effective of time cost using proposed system. As may be seen in Fig.14, for digitalization of 10,000,000 characters, we need 1000 man-days using manual typing method. However, if we use proposed method, we need only 144 man-days.

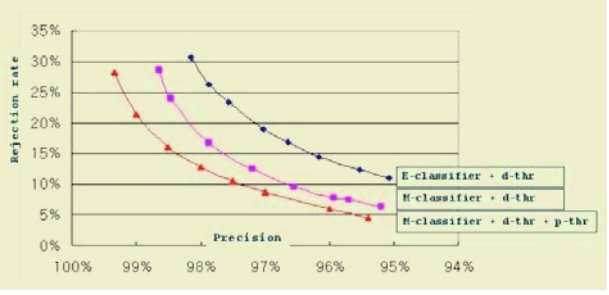

Fig. 13. Comparison of precisions according to rejection rates.

Table 1. Total cost comparison of manual typing only method and proposed method.

Input Type	Total Cost	Input Cost(10 per char)
		Reform Cost(30 per char)
Manual Typing Only	100,000,000	10,000,000×10=100,000,000
		0
Proposed System	18,680,000	1,268,000×10=12,680,000
(Case of precision 0.98)		200,000×30=6,000,000

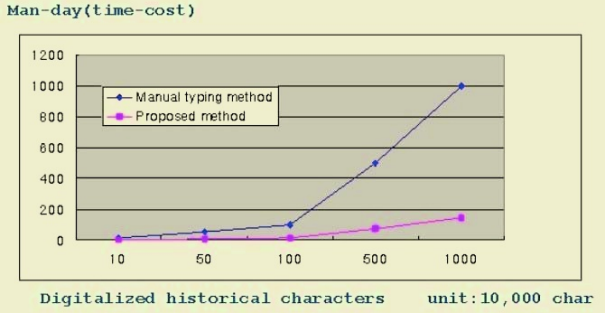

Fig. 14. Time-cost comparison of manual typing method and proposed method.

Assumption :

- a. 1 man-day: characters can be inputted by manual typing for 1 day = 10,000 characters
- b. The number of historical characters must be digitalized = 10,000,000 characters
- c. Accepted ratio = 0.8732
- d. precision = 0.98

Yield:

- Time cost of manual typing method : b/a = 1000 man-days
- Time cost of proposed method: $((b \times c \times (1-d)) + (b \times (1-c))/a = 144$ man-days

5 Conclusions

Just OCR-based Digitalizing of handwritten historical documents is a difficult problem because of complex layouts, blurred characters, shape variations due to writers' habits, styles, and so on. We have been developing a full-fledged system for OCR-based digitalization using a combination scheme between a manual typing and OCR to digitalize historical meterials. In the system, a huge amount of documents were segmented at at once by proposed segmentation method and individual characters were identified with OCR module, and each character with the same class label was collected into one of predefined character groups. After the grouping with OCR, an operator can verify the correctness of the classifications and finally input text codes for each group, instead of typing all the characters. Character segmentation has long been a critical area of the OCR process. In this paper, we proposed dedicated segmentation method uses geometric feature and context information with Viterbi algorithm. Posterior-based Rejection method designed using Mahalanobis distance, too. According to experimental results, we could see proposed methods helped enhancing the overall efficiency of the processs and reducing the costs.

References

1. S. Hara : OCR for CJK classical texts preliminary examination. Proc. Pacific Neighborhood Consortium(PNC) Annual Meeting,Taipei, Taiwan. (2000) 11–17
2. Z. Lixin , D. Ruwei : Off-line handwritten Chinese characterrecognition with nonlinear pre-classification. Proc. Inc. Conf. On Multimodal Interfaces(ICMI 2000). (2000) 473–479
3. M. S. Kim, M. D. Jang, H. I. Choi, T. H. Rhee, J. H. Kim : Digitalizing Scheme of Handwritten Hanja Historical Documents. Proc. Document Image Analysis of Libraries(DIAL2004) Palo Alto, California. (2004) 321–327
4. C. H. Tung, H. J. Lee, J. Y. Tsai : Multi-stage precandidate selection in handwritten Chinese character recognition system. Pattern Recognition. **27(8)** (1994) 1093–1102
5. L.C.Tong, S.L.Tan : Speeding up Chinese character recognition in an automatic document reading system. Pattern Recognition. **31(11)** (1998) 1601–1612
6. Q.Chen , L. Zhen : Word Segmentation in Handwritten Chinese Text Image Based on Component Clustering Techniques. Proc. 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering. **1** (2002) 435–440
7. S.Zhao , Z.Chi , P.Shi , H. Yan : Two-stage segmentation of unconstrained handwritten Chinese characters. Pattern Recognition. **36** (2003) 145–156
8. Y.H.Tseng , H.J.Lee : Recognition-based handwritten Chinese character segmentation using a probabilistic Viterbi algorithm. Pattern Recognition Letters. **20** (1999) 791–806