

Design of a Chinese Name Card Understanding System

Hsi-Jian Lee* and Shan-Hung Lee#

*Department of Medical Informatics

Tzu Chi University, Hualien, Taiwan 970

Email: hjlee@mail.tcu.edu.tw

Department of Computer Science and Information Engineering

National Chiao Tung University, Hsinchu, Taiwan 300

Abstract

In this paper, we present an automatic understanding system for Chinese name cards. After preprocessing of an input card image, we grouped characters into item blocks and segmented characters according to their aspect ratios and gap widths. After character extraction, we sent characters to a statistical multi-font character recognizer. We identified items according to their geometric characteristics and embedded key-characters. A user can edit segmentation results interactively and then save the results in the database for further applications. The high performance in the experiments show the effectiveness of the proposed system.

1. Introduction

Name cards are important communication media to exchange personal information among people. When business activities increase, people will collect more and more name cards. Traditionally, people use card holders to manage these cards. When a person need information of another person, he will find his card manually from the holders. This process is inconvenient and time-consuming. In currently available name card management systems, a user has to input data for each item of a name card through keyboards. This input method is also tedious. In this paper, we aim to design a name card understanding system that will extract and classify the items from card images automatically. Although many papers discuss general document processing, no paper addresses issues about name card processing.

Typically, a document processing system includes the following modules: preprocessing, document layout analysis, character segmentation and grouping, character recognition, and postprocessing [1]. In a card understanding system, preprocessing includes binarization, deskewing and noise reduction. These functions will not be discussed in this paper. The module of document layout analysis finds item blocks from the cards and possibly extracts the logo included. Character segmentation and

grouping decomposes item blocks into individual characters. Character recognition includes feature extraction and comparing the extracted features with those of each reference character. Postprocessing performs content understanding and database construction.

Items in name cards include holder-name, company, title, address, phone-number, fax-number and so on. In general, these items can be printed in different locations according to owner's preference. In other words, a name card has freer layout, while a general document has rather fixed format. Since different items of a name card may have different font types and sizes and an item may have only a few characters, there is less information for character segmentation in name cards than in general documents.

A name card probably has several different fonts. Figure 1 shows a name card, which uses different fonts. Font types in Chinese name cards generally include Ming-Font, Kai-Font and Black-Font. There are significant variations among these fonts, so a general character recognition module may not produce a satisfiable recognition rate for inputs with different fonts. The judgement of fonts used in text lines with only a few characters is not reliable. It is thus necessary to design a character recognition module for multi-font characters.

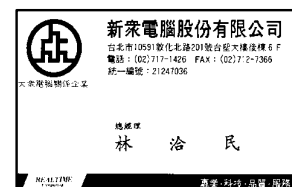


Figure 1. A Chinese name card in different fonts.

A name card may use characters in different languages such as Chinese, English, and numerals. To distinguish Chinese characters with alphanumeric letters, the aspect ratio of a character is usually used. In our system, we divide the recognition module into two sub-modules: Chinese recognizer and English recognizer.

Character segmentation is a technique that partitions an input image into individual characters. Lu [2] presented an overview of machine printed character segmentation. Casey and Lecolinet [3] classified character segmentation methods into three categories: (1) dissection methods, (2) holistic methods, and (3) recognition-based methods. In the first category, text images are cut into meaningful components with predefined properties in character heights and widths, and gaps between characters. Tseng and Chen [4] proposed a method for segmenting handwritten Chinese characters. In the second category, a holistic process recognizes an entire word without segmenting the word. This type of methods is mainly appropriate for alphanumeric characters. Since Chinese characters are non-alphanumeric, it is inappropriate to segment Chinese text into words. In the third category, candidate segmentation positions are detected and then verified by recognizing segmented images. The methods in this category are generally more effective than other methods in segmenting characters in Chinese text.

Techniques used for optical character recognition (OCR) [5, 6] can be roughly grouped into two main categories: structural methods and statistical methods. The former methods represent characters as compositions of structural units, usually called primitives. Techniques such as relaxation and dynamic programming are applied to match an input character with reference characters. The latter ones take characters as whole 2-D patterns, and extract sets of statistical features from the patterns. Since Chinese characters in cursive script have similar and complicated structures, statistical methods are more appropriate than structural ones. In this paper, we adopt a statistical method.

This paper emphasizes on the problems of extracting items from horizontal and vertical cards, character recognition for multiple fonts, and contents understanding.

2. Grouping and Segmentation

Since input name cards are gray-valued and the background levels of different cards are not the same, we first perform a local binarization process to transform the card images to binary images. Then we extract 8-connected-components.

We next decide the image block that a connected-component belongs to. In our system, the task is performed based on geometric properties of card constituents. The connected-components with overlapped bounding rectangles are merged. The merged connected-components are called *character components* hereafter. The character components are sorted based on

the y-coordinates of the connected-components in vertical cards and the x-coordinates in horizontal cards.

For item grouping of vertical cards, the procedure is summarized as follows.

Step 1. For each character component, we find its nearest component group in the vertical direction. If the character component is too far from all existing component groups, a new group is created.

Step 2. Merge component groups with overlapped bounding rectangles.

Step 3. Merge character components overlapped in the horizontal projection profile in each component group as a new character.

Step 4. [Check if there are any component groups to be split.] For the gap between two consecutive character components, if the widths of the vertical projections for the character components upper and below have a significant difference, split the component into two groups from the gap.

After item grouping, we segment characters in each group. The character components are located left to right in horizontal items and top to bottom in vertical ones. For vertical items, the segmentation is performed according to the ratio of the width of a character component to the height its group.

Chinese characters are generally classified as full-characters. If a character is composed of two horizontally aligned sub-characters, this character will probably be classified into two half-characters. In this situation, the multi-font recognition module, to be introduced later, may recognize them as alphanumeric letters or punctuation marks incorrectly. We merge two or three consecutive half-characters if the gaps between them are smaller than the average gap.

Figure 2 shows a character segmentation result. We can see that erroneously split characters are merged correctly. However, two half-characters, numerals “4” and “0”, were merged erroneously since they satisfy the condition above. From the experimental results, we found that this case happens frequently in specific items such as address and phone-number. These items have several key characters and characteristics such as the characters in a phone-number are all numerals. If characters in this item are recognized as Chinese characters, we try to split them and put them into the English recognition module again.

3. Multi-Font Character Recognition

A Chinese name card may use only a single font or several different fonts. Recognizing multi-font characters is not easy, because each font has its own features.



Figure 2. Character segmentation result. Characters marked with * are merged from two half-characters. The last one is merged erroneously.

Due to the large character set, these features may intersect in the feature space. We cannot expect a high recognition rate if we recognize single-font characters with features trained for other fonts.

There are several possible ways to solve the multi-font character recognition problem as listed below.

1. Train the features for all characters in different fonts. There are about 5401 Chinese characters commonly used. For four different commonly-used fonts in Chinese name cards, we have 21604 categories totally. It is not easy to add new features to discriminate these large amount of categories perfectly.
2. Identify the font type of a character first. However, it is not easy to recognize the font type reliably. Especially, when only a few characters are in a specific font, statistical information cannot be trained and utilized very effectively. Tseng and Lee [7] showed that the rate of font identification is only about 82%.
3. Design different character recognition modules for different fonts [8]. Each unknown input character image was fed it into different character recognition modules. Several possible candidates with different fonts were produced. In our system, we adopt this solution due to the following reasons.
 1. We have trained a set of features, including 16-dimensional crossing counts and 256-dimensional contour directional counts, for a specific font of characters. This method can achieve a high recognition rate.
 2. We do not aim to reconstruct the original images of cards. It is not necessary to recognize font types of the characters. We have only to extract items from name cards.

To recognize single font machine-printed characters, we split a character image uniformly. Then we extract the crossing counts and contour direction counts of each character image. To match a character pattern, we calculate the weighted-distance $DIST$ of feature values between an unknown input patten and each Chinese character:

$$DIST = W \times CC_DIST + DIR_DIST$$

where CC_DIST is the distance of crossing counts, DIR_DIST is the distance of contour direction counts,

and W is the weight. In our experiments, we set W to 12 according to the training results.

In the matching stage, we feed each character image into character recognition modules, including machine-printed Chinese character recognizer (CCR) for Chinese characters, and English character recognizer (ECR) for alphanumeric letters and punctuation marks. If both width and height of an input character image are less than 40, the recognition module normalizes it to 40 in width or height depending on the aspect ratio.

If the font type of an input character image is known, the input image will be fed into the relevant recognition module. Otherwise, the input image will be processed first by a clustering algorithm and then recognized by the mixed-font character recognition module [7]. The clustering algorithm reduces the matching time by using the feature of crossing counts. In our experiments, we collect features of four different fonts: Kai-font, Ming-font, Round-font and Li-font. Each font has 5401 character categories, so there are totally 21604 character categories. We first divide all categories into 1000 clusters by means of the nearest-mean reclassification algorithm and then select the top N clusters. The input character image will next match with those characters in the top N clusters. In this way, we can save much matching time.

For the English recognition module, we collect two types of characters, those with serif and without serif. English characters in a vertical name card may rotate 90 degree clockwise. To recognize these rotated characters, we may rotate the images 90 into their normal direction and recognize these character in the ECR. However, it is difficult to measure the rotation angle reliably before character recognition with only very few characters in a card item. Instead, we collect those rotated character samples from training name cards and recognize those rotated character images directly. From outputs of those recognition modules, we choose the best recognition candidate according to a majority voting strategy among the characters in a text line.

4. Item Identification and Post-processing

After character recognition, we will analyze the contents of the cards. Generally, there are several basic items in Chinese name cards, including name, title, company, position, address, phone-number, fax-number and e-mail address. Each item has its own characteristics, to be clarified later.

4.1 Item Identification

For basic items in Chinese name cards, we define basic items from their characteristics and keywords. To

identify items in a Chinese name card, we use the a two-phase procedure. In the first phase, we try to find all possible items except title and education, since the locations of these two items are related to the holder's name. In the second phase, we try to find these two items if they exist.

In phase 1, we find possible items according to the following sequence: name, company, address, phone-number, fax-number, and e-mail address. Under the assumption of a unique holder's name, we first find the name item. Other items are checked according to the order arranged by the frequencies computed from the training samples. We check the characteristics of those items. If a character block satisfies rule preconditions or has key-characters of a specific item, a score will be assigned to the block. The item with the highest score is identified as the candidate item of the input character block.

In phase 2, we first calculate the horizontal or vertical distance between each unidentified item and the name item identified in phase 1, depending on the layout style of a name card. Then we check if there are any blocks satisfying criteria of title or education items. If the answer is positive, the block is assigned as the item satisfying the criteria.

4.2 Candidate re-splitting

After item identification, we use the characteristics of each item to revise character candidates; for example, in the phone-number item, most characters are numerals. If a non-top candidate of a character image is a key-character, we rank it as the top candidate.

Two numerals may be merged together since they can be combined as a full-character. To handle this problem, we perform additional character splitting according to key-characters. We first collect all possible key-characters that have numerals in front of them. When a key-character is found in the candidate list, we check the character candidate in front of the key-character. If the candidate is not a numeral, we try to split it using the projection profile. If there are gaps between the project profile, we split this merged character vertically. After splitting, these characters are recognized by the English recognition module.

4.3 Editing and access

After above procedures, some characters are still wrongly split into two parts or two characters are merged into a character. Through an editing interface, a user can correct the errors by choosing one of three operations: horizontal splitting, vertical splitting and character merging. After the user chooses the operation, the system

sends split or merged characters to the OCR module immediately.

After above processes, card information is stored in the database. The cards with similar attributes can be classified into a category. For example, the cards of friends can be put in the "friend" category. This function makes a user manage his cards easily and efficiently. A user can browse the cards in the database and search the cards through a search engine. The user can find the card information with the card image they need easily. This is one of the primary objectives of the system.

5. Experimental Results and Analysis

Among 590 test name cards, there were 330 horizontal cards and 260 vertical ones. The test card images were scanned at 256 gray-scales with resolution of 300 dpi (dot per inch). We use two criteria to evaluate the performance of item grouping, which are defined as follows:

$$\text{Extraction rate} = \frac{\text{Number of correctly extracted objects}}{\text{Number of total objects}},$$

$$\text{Accuracy} = \frac{\text{Number of correctly extracted objects}}{\text{Number of total extracted objects}},$$

where an object may be an item or a character. Table 1 shows the extraction rate and accuracy. From the table, we can see that the extraction rate is about 95% on average and character recognition accuracy is about 93%.

Table 1. Results of item grouping and character segmentation

		total objects	total extracted objects	total correctly extracted objects	extraction rate	accuracy
Item Grouping	Horizontal	290	297	275	94.83%	92.59%
	Vertical	217	221	205	94.47%	92.76%
character Segmentation	Horizontal	3,421	3507	3277	95.79%	93.44%
	Vertical	2,359	2413	2267	96.10%	93.95%

A text line may consist of Chinese characters, alphanumeric letters and punctuation marks. Since we only use bounding rectangles of characters to perform charac-

ter segmentation, there are probably wrong results. Figure 3 shows several cases. If characters in vertical cards are rotated 90 degree, the shape of the character changes. The segmentation rules cannot work as shown in Figure 3(a). Also, when the key-characters have not been recognized, the numerals might not be split correctly as shown in Figure 3(b).

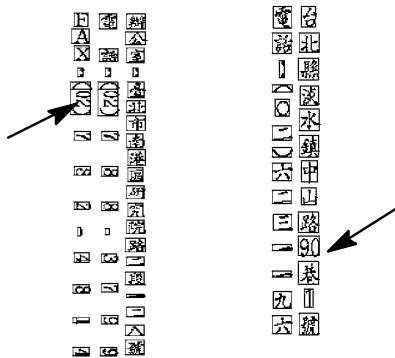


Figure 3. Incorrect results of character segmentation.

In the following, we present the results of multi-font character recognition. Suppose that we select the characters in the top 50 clusters as the character candidates. When the number of clusters is 1000, the recognition rate is the highest (91.68%). With fixed number of clusters, 1000, we check the effect related to the number of candidate clusters selected. We found that the recognition rates do not change significantly after 50 candidate clusters. Therefore, the multi-font recognition module classifies 21,604 characters into 1,000 clusters and set the number of candidate clusters as 50.

The definition of the accuracy for item identification is similar to that of item grouping. The accuracy rate for horizontal cards is 81.54%, and vertical cards 86.62%. In addition to size and layout information, we also apply character recognition results to identify items. However, when the key-characters of an item are not recognized, the item might be classified erroneously. Items having similar characteristics will cause identification difficult. For example, phone-numbers and fax-numbers mainly consist of numerals. If several numerals of a phone number are wrongly recognized as alphabets, those items might be classified as an e-mail address.

6. Conclusions

In this paper, we have designed a Chinese name card understanding system. Our system consists of four phases: preprocessing, extraction of items and characters, character recognition and item identification. In the

first phase, we applied the dynamic thresholding method to separate the foreground from the background. Then we extracted connected-components and deleted noises, marks and lines. In the second phase, we extracted item blocks line by line. In the third phase, we sent the character images into the recognition modules. The recognition modules include mixed-font Chinese character recognizer, specific-font Chinese character recognizer and English character recognizer for alphanumeric letters and punctuation marks. In the last phase, we identified card items according to their characteristics and key-characters.

Several suggestions for improving the proposed system are listed as follows.

1. The system should be extended to process cards with color-inverse items and with non-uniform background.
2. Some items are broken into two or more parts. The connection of these items to obtain a complete item is left for future study.

7. References

- [1] S. Tsujimoto and H. Asada, "Major Components of a Complete Text Reading System," *Proc. IEEE*, Vol. 80, No. 7, July 1992, pp.1133-1149.
- [2] Y. Lu, Machine printed character segmentation - an overview, *Pattern Recognition* 28 (1996), 67-80.
- [3] R.G. Casey, E. Lecolinet, A survey of methods and strategies in character segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.* 18 (1996), 690-706.
- [4] L.Y. Tseng, R.C. Chen, Segmenting handwritten Chinese characters based on heuristic merging of stroke bounding boxes and dynamic programming. *Pattern Recognition Letters* 19 (1998) 963-973.
- [5] L. Tu, et al., Recognition of handprinted Chinese characters by feature matching, *Int. Conf. on Computer Processing of Chinese and Oriental Languages* (1991) 154-157.
- [6] H. J. Lee, Chinese character recognition in Taiwan, in: *Handbook on Optical Character Recognition and Document Image Analysis*, World Scientific Pub., 1996, pp. 331-355.
- [7] Y.H. Tseng, C.C. Kuo, H.J. Lee, Speeding-up character recognition and in an automatic document reading system. *Pattern Recognition* 31 (1998) 1601-1612.
- [8] L. Xu, A. Krzyzak, and C.Y. Suen, Methods of combining multiple classifiers and their applications to handwriting recognition, *IEEE Trans. System Man Cybernt.* 22 (1992) 418-435.
- [9] C. H. Tung, *A Study of Handwritten Chinese Text Recognition*, Ph.D. Dissertation, Institute of Computer Science and Information Engineering, National Chiao Tung University, Taiwan, R.O.C., 1994.