UNIVERSITY OF CINCINNATI

October 25, 20 02

I, Marie N. HU,
hereby submit this as part of the requirements for the degree of:

Master Of Science

in:

College of Engineering

It is entitled:

A Study Of Chinese Character Recognition Methods

Approved by:

Dr. Anca Ralescu

Dr. Ken Berman

Dr. Carla Purdy

A Study of Chinese Characters Recognition Methods

A thesis submitted to the

Division of Graduate Studies and Research

of the University of Cincinnati

in partial fulfillment of the

requirements for the degree of

Master of Science

In the

Department of Electrical and Computer Engineering

And Computer Science

Of the College of Engineering

August 2002

By

Marie HU

B.S., University of Cincinnati, 1997.

Master of Advisor in Computer Business Application(DESU), University Paris VIII (France), 1987.

PhD in Doctorat of Third Cycle of Social Study, Institut d'Etudes Politiques de Paris (France),1983.

Thesis Advisor and Committee Chair: Dr. Anca Ralescu

Abstract

The use of the tablet and pen to input the Chinese Characters into computer is no longer a dream. During the last decade, different researchers have tried to find the solution to this problem. Although some results have been achieved, the use of tablet and pen is still not wide spread.

In the introduction, I explain the Chinese character elements and their structural composition. Through different methods applied in Chapter 2 of the thesis paper and the presentation of Handwritten Chinese Character Recognition(HCCR)problems, I present the fuzzy and structural approaches in the preclassification phase. The task of connecting the tablet and pen to the recognition of final Chinese character is considered in system design.

Due to the complexities of the HCCR problems, this thesis research paper will first concentrate on the methodology by exposing theoretical system slection. My objective is to propose a practical, simple, and feasible method of HCCR and its preliminary implementation for the beginner interested in this problem. The algorithms for stroke extraction by finite automata and character matching by Depth First Searching (DFS) require specific conditions for preparing Dynamic Programming (DP).

The next step will be setting up several databases, finding the appropriate tool of the tablet and pen, connecting correctly the tool to the computer, and finishing the postprocessing phase. These tasks are beyond the scope of this thesis.

Contents

1 Introduction 6

1.1 Chinese Orthography 7

1.1.1 Elements of a character 9

1.2 The difficulties in Handwritten Chinese Characters Recognition(HCCR) . 13

1.3 The general problems concerning Chinese Characters: 14

1.4 The difficulties in Chinese Characters major recognition process 15

2 Fuzzy Rules-Based method with radical extraction 17

2.1 First use of Fuzzy logic for CHCCR in Taiwan 17

2.2 Other Tendency of CHCCR Fuzzy Methodology in Taiwan 18

2.3 Fuzzy rule-based method with radical extraction 19

2.3.1 System architecture 19

2.3.2 Stroke extracting strategy 19

2.3.3 Primitive strokes 21

2.3.4 The range for first stroke types 21

2.3.5 Regularity degree 21

2.3.6 Measurement on stroke distance and the calculations 22

2.3.7 Fuzzy rules 22

2.3.8 The twenty most frequently used radicals 22

3 **Overview of proposed approach in preclassification phase** **24**

3.1 The Proposed Method: 24

3.1.1 A general block diagram of the Chinese Character on-line recognition system 25

3.1.2 Presentation of whole real time CHCCR on-line process: 27

3.1.3 The connection between the tablet and pen with the computer: . 28

3.1.4 More detailed block diagram of the Chinese Character on-line recognition system 29

3.2 Structural method: 30

3.2.1 The importance of structural method: 30

3.2.2 The detail of structural method 31

3.2.3 Concepts and system overview of structural method. 32

3.3 Shape classification. 38

3.3.1 Locus direction: locus means the exact place of stroke 39

4 **Stroke Number free and stroke Order free recognition** **40**

4.1 Motivation for a stroke-number-free and stroke-order-free for basic stroke recognition: 40

4.2 Objectives to reach in the current implementation: 42

4.3 Summary description of methodology in basic stroke recognition: 42

4.4 Detailed explanation about this proposed method 43

4.4.1 Key features of methodology 44

4.4.2 Geometric feature measures 45

4.4.3 The Current Method 48

4.4.4 The step of stroke correspondence 49

4.4.5 Types of stroke matching 50

4.4.6 Distance Functions 51

4.5 One simple algorithm for stroke extraction by finite automata 52

4.6 Another simple algorithm for character matching by DFS 53

4.7 Decomposition of radicals 56

5 Discussion of the contribution of approach and conclusion 57

6 Future Work:Data Type Header File and 4 Chinese Characters 61

List of Figures

1.1 Basic strokes and their variations[RB80] 10

1.2 Demonstration of sequence of writing[RB80] 12

2.1 System Architecture of Fuzzy Rule-based Method[HHS98] 20

2.2 The 20 most frequently used radicals[HHS98] 23

3.1 Concepts about Chinese Character on-line recognition system between tablet and software[A.N01] 26

3.2 Ten Patterns For Chinese Character Radicals[WFW97] 34

4.1 The eight Reference Point of the bounding Rectangle of a Chinese Character.B:Bottom C:Center T:Top R:Right L:Left 46

6.1 Chinese character:Europe 68

6.2 Chinese character:Talkative 68

6.3 Chinese character:Lake[WFW97] 68

6.4 Chinese character:Address[WFW97] 69

List of Tables

3.1 Locus Direction Table 39

4.1 Minimum Bounding Rectangle Table 47

Chapter 1

Introduction

The Chinese language is conceptually very different from Western languages. In addition to there being thousands of Chinese characters to recognize, compared to only a few tens of characters from Western languages, they can also be read and written from left to right, right to left or top to bottom with any of these styles intermixed within a single document. These differences have an impact on the on-line Optical Character Recognition (OCR) and off-line Handwritten Chinese Characters Recognition (HCCR).

In Chapter 1 I will examine Chinese Orthography according to the following criteria: its elements of a character, basic strokes in character formation and radicals.

According to Tang Yuan Y [TT98], Chinese characters have three characteristics:

- The vocabulary is typically large, involving more than 3,000 classes.
- The structures of Chinese characters are much more complex than those of alphanumeric letters.
- Many Chinese characters have similar shapes.

According to ATIP (Asian Technology Information Program), an autonomous US non-profit organization that collaborates with other public, private, and educational organizations worldwide to analyze and disseminate information about Asian Science and Technology), Chinese is arguably better suited for handwriting than for printing or typing. It is impossible to fit the many thousands of distinct Chinese characters onto a reasonable keyboard. It is astonishing to realize that Chinese and Japanese business development as late as the 1980's relied on the pen as the primary tool of written personal and business communication.

However, recently the widespread use of personal computers has changed this picture. Of the three methods commonly used to enter Chinese characters into a computer, it seems the most natural is on-line OCR. But, the Chinese handwritten recognition is a most difficult task, for which, at that time, no perfect method was discovered.

1.1 Chinese Orthography

According to P. Wang [Wan01], the earliest Chinese ideographs were found on shells and bones dating back some 3,400 years. Judging by the extent to which these shell-and-bone logographs are already conventionalized according to the rules, it is reasonable to infer that true writing emerged considerably earlier, although we do not know exactly when.

The earliest Chinese language dictionary is Shuowen Jiezi (121 A.D.), a compilation of 9,353 logographs. Its author, Xu Shen, applied one principle of organization that was accepted and used by all schools: the logographs was divided into six categories (methodologies) according to the way they were formed:

- **Pictographs(Xinagxing)** are iconic and imitative drafts. For example, the character for the word mountain looks just like a mountain and that for horse is complete with four legs.

- **Simple ideograms(Zhishi)** are formed with pictographs to suggest an idea (indicative letters).

- **Complex ideograms(Huiyi)** are logical aggregates that follow logical reasoning.
- **Phonetic loans(Jiajie)** are borrowed words.
- **Phonograms(Xingsheng)** are made up of two or more components (phonetic complex). While the signific suggests its meaning, the phonetic indicates its pronunciation. Phonograms are the most productive category, estimated in a Qing dynasty (1644-1911 A.D.) study to comprise 82%, and in a later study about 90% of the logographs read.

  In this section, I focus my attention to Chinese orthography of the phonograms. Logographs are identifying symbols showing a system of relationships between things. Phonograms are generally formed from two components: a component that suggests the meaning, the other that indicates the pronunciation. The category of phonograms is the most important for the following reasons:

  1. The majority of the logographs are of this type (about 90%).
  2. This ingenious scheme simultaneously presents cue sounds and meaning, providing the most rational foundation for most orthographies.

- **Derivatives(Zhuanzhu)** are derived from generalized words.

  Although Chinese orthography has had a long and continuous development, there were only two great epochs in the history of the Chinese characters [Wan01]:

  1. From earliest times to the Qin Dynasty.
  2. From the Small Seal(XiaoZain period) of 200 B.C to the present: the modern script is based directly upon Small Seal style that practised at Small Seal period.

According to P. Wang [Wan01], Chinese characters are two-dimensional abstract images, or patterns, which are the ideal vehicle for investigating problems related to pattern recognition, abstraction, and communication of information.

1.1.1 Elements of a character

The elements of Chinese characters may be viewed at two distinctive levels:

A. Strokes: it refers to the most basic elements.

B. Radicals: it refers to the structural parts. The radicals are of different sizes and have to be placed in specific positions with respect to each other to make a legitimate character.

In the following section, we will explain these two elements in detail.

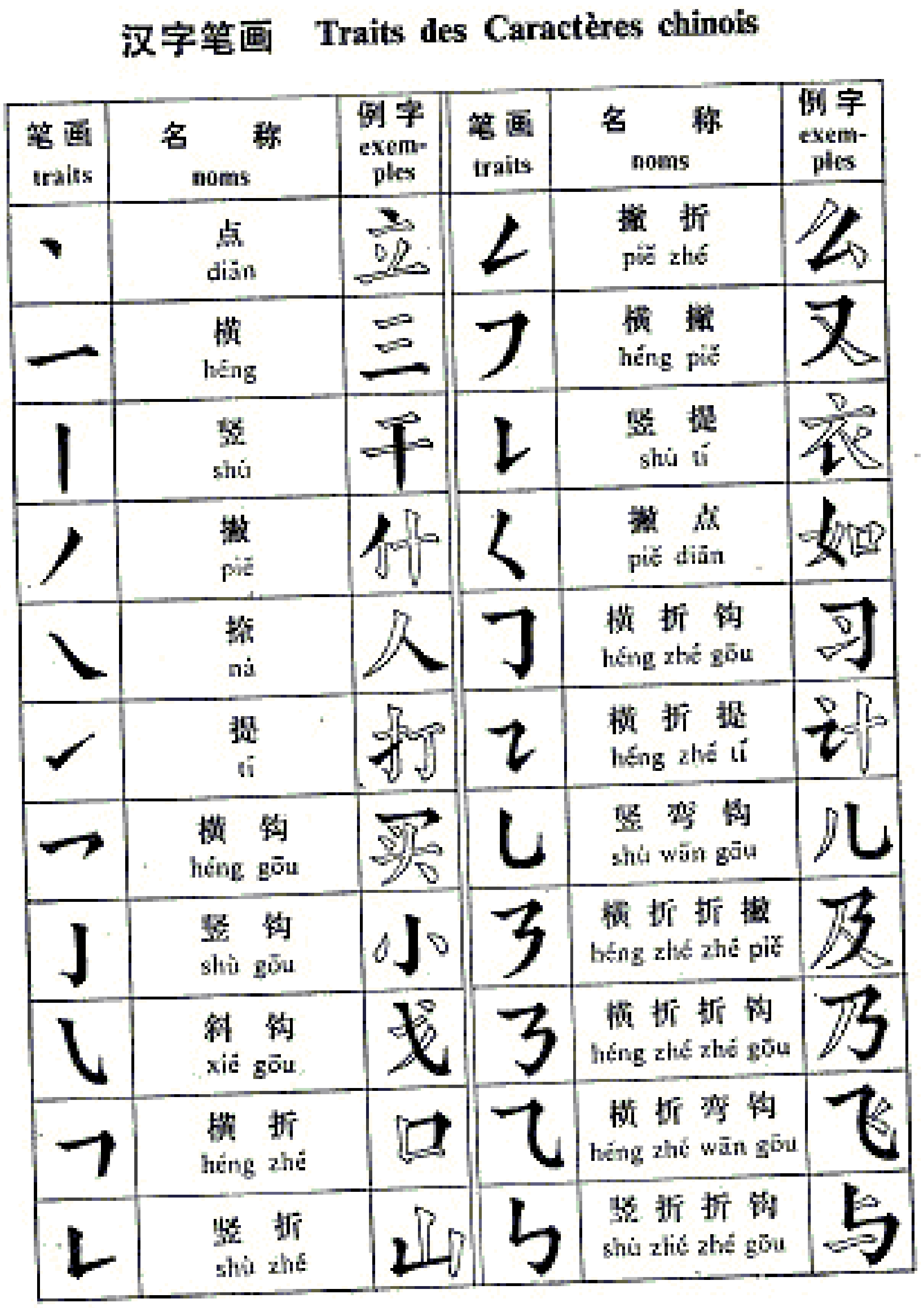
A. Basic strokes in Chinese character formation.

Calligraphy, the elegant rendering of characters, is a highly cultivated art form long prized in Chinese culture. For the Chinese, a harmonious relation exists between painting and calligraphy.

From the calligraphic standpoint, a character is able to be reduced(reducible) to simple strokes, which are the material elements of modern writing. According to P. Wang [Wan01], there are nine basic strokes. But if we include variations, then it is increased to seventeen. However, according to my observation, this is not the same for the variations, and in fact, there are more than sixteen variations.

According to the Beijing Radio Broadcasting of Chinese Language Book [RB80] the basic strokes are the following:Dian, Hang, Shun, Pie, Na, Ti, Gou, Zhe, and Wang. The variations are Heng gou, Shu gou, Xie gou, Heng zhe, Shu zhe, Pie zhe, Heng pie, Shu ti, Pie dian, Heng zhe go, Heng zhe ti, Shu wang gou, Heng zhe zhe pie, Heng zhe zhe gou, Heng zhe wang gou, and Shu zhe zhe gou.

Chinese Characters are written according to the following rules (see Fig.1.2):



汉字笔画 Traits des Caractères chinois

笔画 traits	名称 noms	例字 exemples	笔画 traits	名称 noms	例字 exemples
丶	点 diān	立	㇜	撇折 piě zhé	么
一	横 héng	三	㇇	横撇 héng piě	又
丨	竖 shù	千	㇙	竖提 shù tí	衣
丿	撇 piě	什	㇛	撇点 piě diān	如
㇏	捺 nà	人	㇆	横折钩 héng zhé gōu	习
㇀	提 tí	打	㇊	横折提 héng zhé tí	计
㇖	横钩 héng gōu	买	㇟	竖弯钩 shù wān gōu	儿
亅	竖钩 shù gōu	小	㇋	横折折撇 héng zhé zhé piě	及
㇂	斜钩 xié gōu	戈	㇌	横折折钩 héng zhé zhé gōu	乃
㇕	横折 héng zhé	口	㇈	横折弯钩 héng zhé wān gōu	飞
㇗	竖折 shù zhé	山	㇉	竖折折钩 shù zhé zhé gōu	与

Figure 1.1: Basic strokes and their variations[RB80]

1) Top-down diretion:First horizontal then vertical.

2) Left-right:First left then right.

3) From upper to down part.

4) From left to right.

5) From outside to inside.

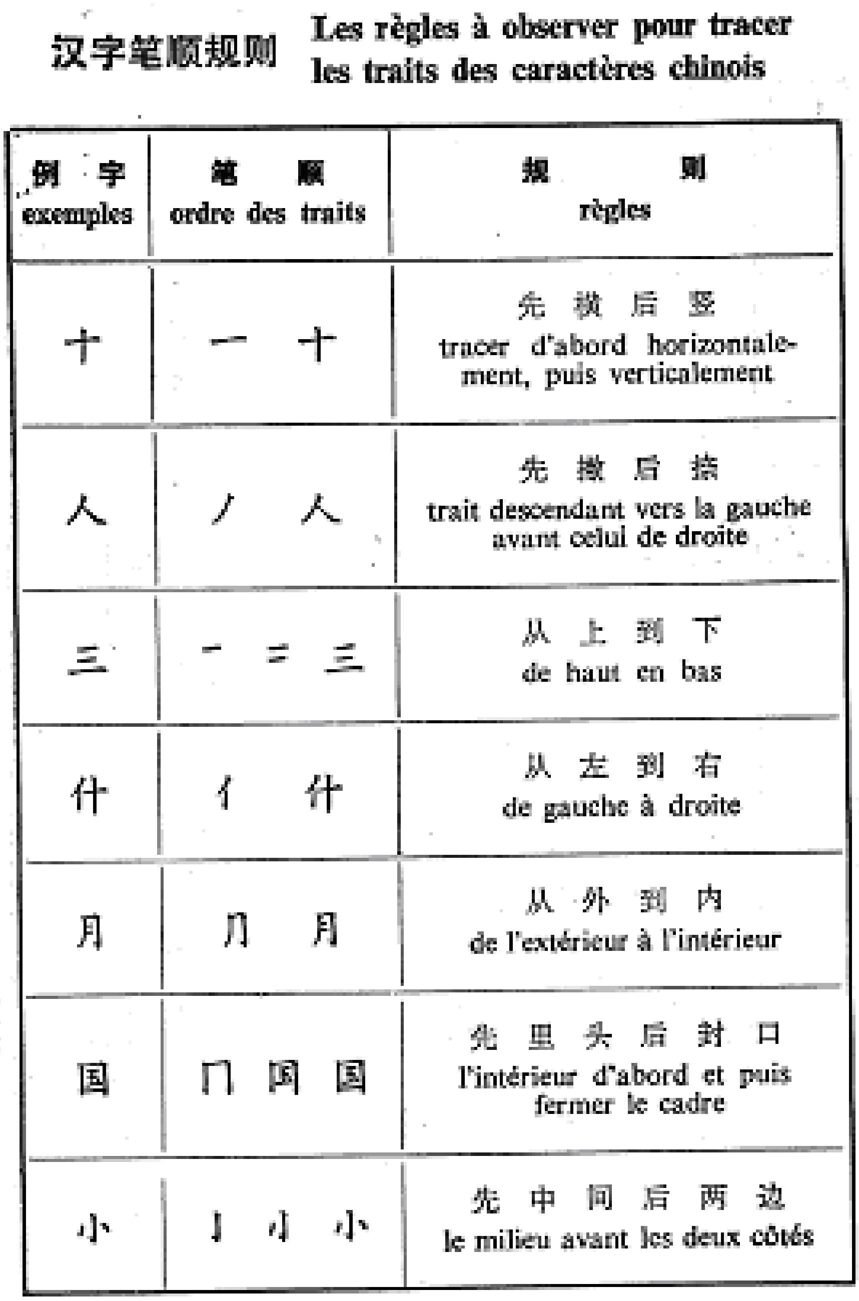
6) First inside content then close the box.

7) Middle is completed before the outside: First middle then both sides.

Figure 1.2 demonstrates the sequence of writing by illustrating these rules by showing the sequence of writing some Chinese words. The sequence of writing is very important to learn. At the primary school, the students have to learn the sequence of writing starting in first grade class. It is the same to play the piano. Without correct sequence, writing would be like playing the piano without following the correct finger steps. The non-respect of sequence will have serious consequences (effects) to the final results. A Chinese character cannot be drawn, it must be written according to the rules mentioned above. Failure to do so results in poor Characters and it is unacceptable.

Moreover, according to P. Wang [Wan01], the teacher always makes a big point about the proper stroke order of writing. It is not uncommon that in teaching and learning a character, it is copied fifty or sixty times, or even more. But, of the millions of Chinese children who begin the learning process, only a small percentage ever arrive at stage of “wen-li tung” which practically means to be able to read and write literature and composition intelligently.

According to P.Wang [Wan01], the basic units are a very small number of distinctive items. The beauty of simplicity makes it easy to remember and to work with for a beginner. On the other hand, being relatively few in quantity, the necessity appear in different characters more often reducing their effectiveness as distinctive and specific symbols. Redundancy is an inherent problem when working at the stroke level.



汉字笔顺规则 Les règles à observer pour tracer les traits des caractères chinois

例字 exemples	笔顺 ordre des traits	规则 règles
十	一 十	先横后竖 tracer d'abord horizontalement, puis verticalement
人	丿 人	先撇后捺 trait descendant vers la gauche avant celui de droite
三	一 二 三	从上到下 de haut en bas
什	亻 什	从左到右 de gauche à droite
月	冂 月	从外到内 de l'extérieur à l'intérieur
国	冂 国 国	先里头后封口 l'intérieur d'abord et puis fermer le cadre
小	亅 小 小	先中间后两边 le milieu avant les deux côtés

Figure 1.2: Demonstration of sequence of writing[RB80]

During the 1980's, many researchers studied basic strokes recognitions. The strokes are important and very useful to the beginners who can only recognize them in a new character. On this ground alone, strokes fulfill a need. There are many existing dictionaries that use index systems employing stroke counts as a means of locating a character besides the radicals index system.

B.Radicals.

Identification of radicals in characters is also difficult. Several researchers have focused on this topic [LW98] [YF96] [LN93] [WFW97] [JZ99] [SH84]. To use a Chinese Dictionary one must know radicals. Radicals are the index system in the Western Dictionary.

Some words are also radicals in Chinese Characters such as Man, Mouth, and Hand. For this reason, when a computational analysis of the structural compositions frequently used Chinese characters is done, it is difficult to separate the strokes and radicals. For this important reason, I concentrate my study in this radical's field besides strokes field.

1.2 The difficulties in Handwritten Chinese Characters Recognition(HCCR)

Several difficulties are associated with the task of HCCR are as follows:

- **Word versus Chinese Character**:In many cases, a single Chinese Character represents a word. But very often words are formed by combining different Chinese Character.

- **Complex structure**:Most Chinese Character consist of sub-patterns or radicals. More than two hundreds different radicals exist. A radical can itself be a Chinese Character.

- **Segmentation**:Each Chinese Character is usually separated spatially. There is no need for segmentation between Chinese Character. But in the recognition process, we still need to have segmentation inside one Chinese Character. An exception is artistic writing, in which a whole line of characters can be written using only one component. In order to find the right radical inside one Chinese Character, the segmentation inside one Chinese Character should be handled carefully.

- **Evolution of Chinese Character**: While Chinese Characters in Taiwan and Hong-Kong remain almost as they were 50 years ago, many Chinese Characters in the mainland China have been simplified since the 1950s. This further complicates the character alphabet.

- **Writing style**: Different persons have very different handwritten styles.

  In block or printing style, a character is written component by component, radical by radical, all within a character box. In contrast, in cursive style there is no constraint on the number of components. Several block style components may be connected to form a single component. On average, a Chinese Character has 8 to 10 components. The simplest Chinese Character has 1 component while the most complicated has more than 30 components.

1.3 The general problems concerning Chinese Characters:

Difficulties arise in the study of Chinese Characters from the fact that although many different methods of recognition have been adopted, none yields perfect results. Besides what I have already mentioned in this section, I will restate the general problems concerning Chinese Characters themselves [CLC88].

- **Variation in characters according to individual characteristics of the writer**: The writing of characters may vary from person to person.

- **Variation within radical pairs**: Some radical pairs are written differently.

- **Complexity of Chinese Characters**: Many Chinese characters are very complicated without unique writing sequence and they also could be written differently or incorrectly.

- **Variation in dot type**: The dot type stroke (') may be written differently. It may be identified as different strokes on different occasions.

- **Variation of the stroke count**: The stroke count is also unstable since a character may be mistakenly written in different ways.

- **Existence of characters with different meaning and appearance**: There are many pairs of Chinese Character which consist of the same set of strokes. Many pairs even have the same stroke sequence.

1.4 The difficulties in Chinese Characters major recognition process

Tappert [TSW90] (p790) explained the recognition problems for handwriting and drawing on tablets as following:

- **Recognition of language symbols**: large alphabet of Chinese characters

- **Equations**: inappropriate mathematical equations

- **Line drawings**

Various tasks are associated with the recognition process steps, as follows:

- In preprocessing of handwritten data, the tasks are noise reduction, smoothing, thinning /filtering.

- In normalization process, the tasks aim at correcting anomalies.

- In shape recognition process, the methods for characters, cursive script, words, gestures, equations, line drawings, and signatures are used and discussed. The purpose of shape recognition is the pattern recognition of shapes of writing units. Different approaches to achieve this can be used:

  1. Some shape recognition methods rely on prior analysis of the characters. Features and sequences of code zones can be alphabet specific. The time sequence of zones, directions, or extremes methods rely primarily on dynamic information. Chain Codes method (using code of connection) are used.

  2. Another approach is analysis-by-synthesis, or recognition by generation. Related to this approach is a theory of handwriting perception in which the dynamic information is inferred from the static form: pairwise distinction or by the number, order, and relative position of strokes.

  3. A similar approach used dynamic programming to match real and modeled strokes.

- In the postprocessing process, the output from shape recognition is processing. A postprocessor can operate on this information to obtain estimates for larger linguistic units such as words. Several methods produce a list of words in order of decreasing likelihood according to shape recognition scores. Subsequent dictionary lookup can choose the dictionary entry with the best shape recognition score. Hypothesis generation and test is a common approach. Higher levels of linguistic rules such as syntax and semantics can also increase the recognition rate.

Chapter 2

Fuzzy Rules-Based method with radical extraction

2.1 First use of Fuzzy logic for CHCCR in Taiwan

In 1988 several researchers at Taiwan Tsing Hua University already tried to use the fuzzy approach to solve the CHCCR problems [CHC89]. They found this method had several advantages:

- No distribution of information is needed. - It provides a set of operations for the inference for the property of a fuzzy set.
- It needs less computing time with recognition rate of 96%.

Their method used the following fuzzy concepts:

1. Chinese character could be viewed as a collection of line segments called strokes.
2. Two membership functions are defined for the location measure and type measure between two strokes.
3. A function of fuzzy entropy is used in information measure.

4. They used the modified Hungarian method that related to maximum perfect matching a weighted bipartite graph to resolve the assignment problem.

5. They considered the fuzzy approach to be a useful tool for the OCR researchers.

2.2 Other Tendency of CHCCR Fuzzy Methodology in Taiwan

Because of the efficient and promising result of fuzzy Rule-Based method in 1997, the same researchers continued to apply the fuzzy set theory. In 1998 they created other primitive and Fuzzy features via Neural Net Model. They found that the fuzzy min-max neural network is verified to be superior to two traditional statistical classifiers such as the nearest-neighbor and the minimum-mean distance classifiers.

In 1995 other researchers such as D.C. Tseng, H.P. Chiu, and J.C Cheng have already used fuzzy ring data to do invariant handwritten Chinese Character recognition [CT97] with a scanner. They discovered that the use of invariant features is superior to that of moment invariants.

Recognizing handwritten Chinese words is very difficult due to high complexity and variability of Chinese character. When added to the requirements for rotation-invariant recognition, the problem becomes even more difficult [SH98]. While in Australia in 1997, H. Yan and P.N. Suganthan proposed a fuzzy attribute graph using an OCR scanner that included numerical, vector relational, and symbolic relation attributes:

- Numerical attributes: Length of stroke and distance between various point on the strokes.

- Vector relational attributes: Translation, scale, and rotation of strokes.

- Symbolic relation attributes(stroke types, intersection types.)

The Chinese characters are represented by attributed relational graphs (ARG) using strokes as ARG vertices. A number of vector relational attributes are also used in the representation to improve the performance of the translation and scale invariant as well as rotation sensitive recognition system. It is a very useful study of CHCCR by constrained graph matching.

2.3 Fuzzy rule-based method with radical extraction

I adopted a simple, meaningful, and useful design based on radical extraction which takes advantage of fuzzy set theory and meets the needs of dealing with Chinese character whose writings are fuzzy in nature. The advantages of this approach include the fact that a few fuzzy rules are required. Few hierarchical rule sets are used. The computation effort is not difficult.The capability of the whole system can be enhanced by increasing the number of fuzzy rules appropriately. This method has great flexibility.

Based on the literature reviewed, I want to creat one simple but complete method of classification and to reach my goal of designing one useful HCCR system. This original fuzzy rule-based method has an optimistic 99.63% rate of divisible Chinese character recognition. HCCRBASE provided by computer Communication Laboratories(IRTI) is used to verify the feasibility of this method. [HHS98]

2.3.1 System architecture

The system architecture can be illustrated as Figure 2.1 with last level recognition resulting step.

2.3.2 Stroke extracting strategy

Stroke extracting strategy uses a two-pass fast extraction strategy:

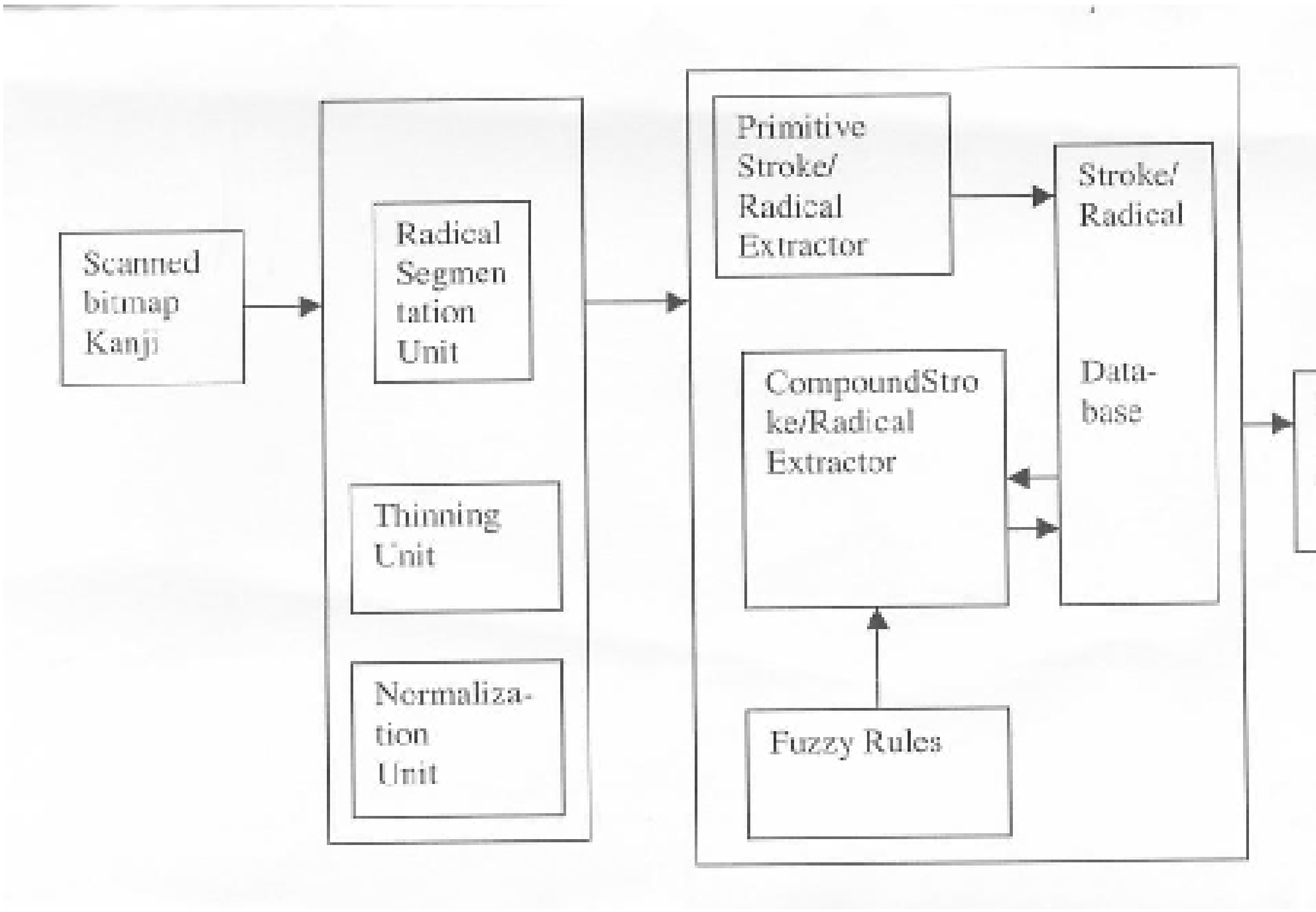


Figure 2.1: System Architecture of Fuzzy Rule-based Method[HHS98]

- Pass 1 checks the Chinese character image row by row and extracts out vertical strokes and slanting strokes
- In Pass 2, the Chinese character image is examined by column-oriented tracing approach. The horizontal strokes are extracted.

A combining phase then proceeds to combine the strokes into corner strokes: A corner strok is composed of two primitive strokes. It also has a direction change in its stroke contour. Finally, it outputs the extraction results.

2.3.3 Primitive strokes

There are seven primitive strokes: HL(horizontal), VL(vertical), LS(left slanting), RS(right slanting), RC(right corner), LC(left corner) and BC(Bottom corner).

2.3.4 The range for first stroke types

If one stroke has been found in the shaded area, it gets the chance to be a candidate of that stroke type.

2.3.5 Regularity degree

To commit/adjust the fuzzy characteristics, a special $\alpha - level\ cut$

degree of regularity is used to determine the regularity of handwritten characters.

If the degree of the extracted primitive stroke for one specific stroke type does not meet the requirement of alpha level, it would not be considered as candidate of this stroke type.

Under the adjustment of regularity, the number of candidates of stroke-type provision can be controlled easily. Since fuzzy rules will exclude impossible stroke or radical combinations, lower REGULARITY is preferred.

2.3.6 Measurement on stroke distance and the calculations

There are mainly three connection types for Stroke 1 and Stroke 2:

- Cross: S1 is crossing the S2. Example such as radical (Ten).

- Meet: S1 only meets the S2. Example such as radical (Man).

- Apart: S1 and S2 are separately found. Such as radical (Water):

Once the connection type of two consecutive strokes is determined, fuzzy rules that represent the character structure will be applied to combine them.

I will use a combination degree for a compound stroke:

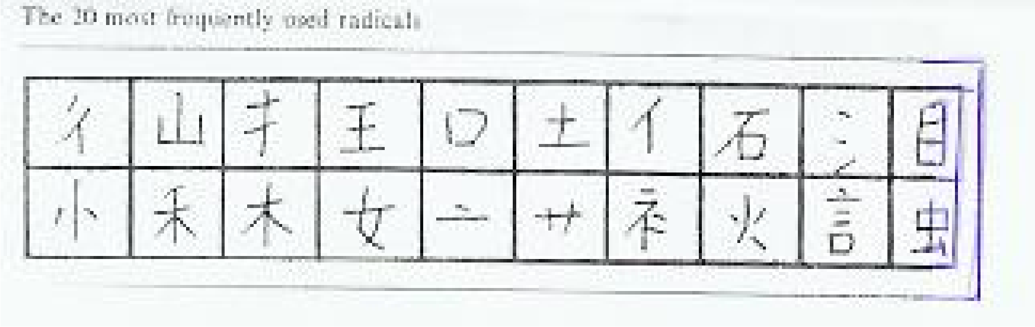
Combination Degree(S1, S2) = min(deg(S1), deg(S2)) where deg() denotes the stroke type degree of the specified stroke S1 and S2.

2.3.7 Fuzzy rules

The different kind fuzzy rules about structure of a radical will be set up in the extractor. These fuzzy rules will include all possible radicals. I can modify easily all rules in one centered site.

2.3.8 The twenty most frequently used radicals

The Figure 2.2 shows the twenty most frequently used radicals.



The 20 most frequently used radicals

彳	山	扌	王	口	土	亻	石	氵	日
忄	禾	木	女	亠	艹	礻	火	言	虫

Figure 2.2: The 20 most frequently used radicals[HHS98]

Chapter 3

Overview of proposed approach in preclassification phase

To illustrate the approach proposed in this thesis, my attention is focused on the analysis of the stroke at the preclassification stage.

3.1 The Proposed Method:

My proposed approach is based on the structural approach combined with fuzzy method in order to resolve the Chinese Character classification problem. The two approaches are used as follows:

- **Structural Approach**: For type and position classification and radical classification.

- **Fuzzy Approach**: For total general HCCR process system (see Chapter 2).

Basic stroke recognition is a minor problem compared to other problems in Chinese character recognition. In reality, the basic stroke recognition really depends on the methodology followed, which regardless of the methodology used, is finally established

. I do not think this causes the problem because basic stroke recognition has existed a long time ago. When the major system methodology has been decided, this basic stroke recognition only needs a programmer to do the coding work.

In this chapter, the following concepts are proposed for design and implementation:

- Chinese character on-line recognition system, Figure 2.2 illustrates the general block diagram for the system.
- Enhanced structural stroke analysis for shape and position classification.
- Implementation. A simple stroke number and stroke order free method for basic primitive stroke recognition.

3.1.1 A general block diagram of the Chinese Character on-line recognition system

.

The recognition system described here includes a tablet and pen interfacing with a host Personal Computer that includes a display screen and a keyboard. While a user writes on the tablet using pen, a recognition program which is running on the PC (recognizer) recognizes the individual block characters written by the user and display both the handwritten Chinese Character and the already recognized character on displayed screen.

The following concepts and terminology are used to describe the operation of the recognizer:

- Touching the stylus to the tablet initiates a *pen down event*, which continues until the stylus is lifted off the tablet for the first time. As soon as the pen is lifted off the tablet, the *pen down* event terminates.

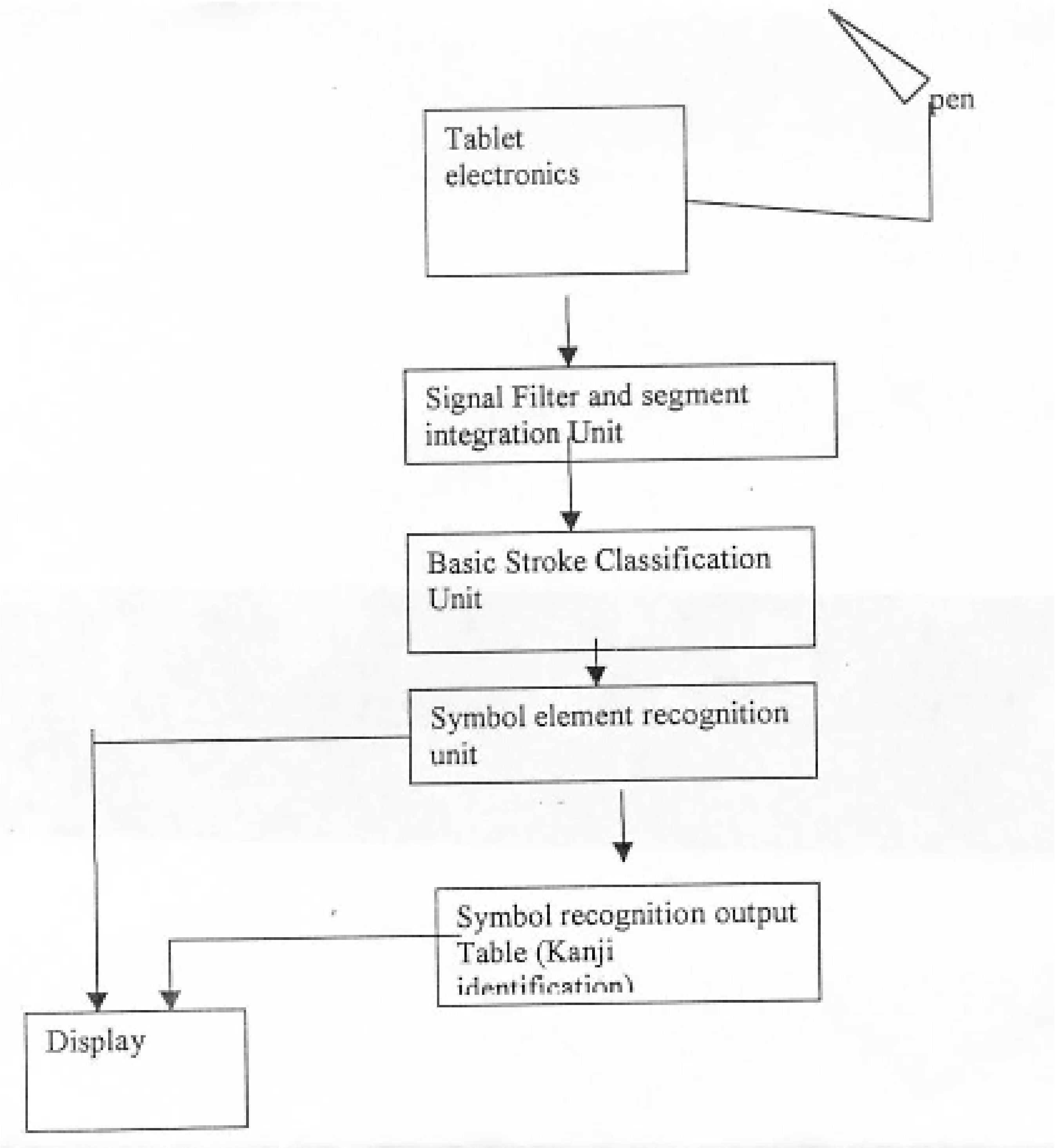

pen
Tablet electronics
Signal Filter and segment integration Unit
Basic Stroke Classification Unit
Symbol element recognition unit
Symbol recognition output Table (Kanji identification)
Display

Figure 3.1: Concepts about Chinese Character on-line recognition system between tablet and software[A.N01]

- The *stroke* is a set of the points representing the sampled pen positions describing the user's input during the *pen down event*. A *stroke* is typically made up of segments, each of which is an object that is associated with part of the input image obtained while the pen is traveling in one direction. A *stroke* commonly changes direction more than once. Therefore, a stroke often includes more than one segment.

3.1.2 Presentation of whole real time CHCCR on-line process:

The whole real time CHCCR on-line process is my first subject of presentation in this paper:

- A handwritten character recognizer has an input cluster buffer and a point buffer with feature extraction and segment analysis by conical boundaries for identification of stroke segments with stroke extractor.

- A stroke recognizer compares single copies of idealized stroke representation with hierarchically approximated multiple scaled topological representations of a current stroke.

- Compared a selected topological representation of the current stroke with boundaries defined by linear combinations of features of direct and reversed ideal stroke prototypes to provide a stroke identification.

- A cluster recognizer maintains a time ordered current stroke buffer and previous stroke buffer and constructs a per stroke area of influence list.

- The time ordered buffers are scanned to generate a spatially ordered buffer.

- A position discriminator assigns character meanings to clusters of strokes in the same window buffer.

- The buffer scanner is responsive to a current stroke for reordering the toke buffers and determining new cluster meanings.

- An editor is responsive to an editing gesture or combination of two strokes for directing the stroke buffer controller to modify the strokes in the stroke buffer accordingly.

3.1.3 The connection between the tablet and pen with the computer:

In his article "Real Handwriting Recognition System", N.A.Jourjine mentioned the importance of connection between the tablet and pen with the computer.[A.N01]

A. In the Data Type header file I propose the following program in C (Chapter 6)for the description of different data structures with data types which will be done prior to basic strokes recognition.

B. All of these connections are intended to demonstrate that without using correct tablet and pen procedures, we could not extract the basic strokes even if the basic strokes extraction method is correct. These connections between different tools are very important to prepare. As I mentioned in footnote 1 at the end of Chapter 2, in Taiwan the basic stroke extraction method was invented a long time ago by several universities including Chiao-Tung (see note at the end of this chapter). Several researchers continued to use this basic strokes laboratory for further advanced researches in CHCCR. [AKF96] [LJCL90].

3.1.4 More detailed block diagram of the Chinese Character on-line recognition system

In order to advance in my research, I contacted the Waco Company that specialized in pen and tablet hardware.

- After contacting the Waco Company in USA that specialized in the field, I realized the tablet electronics provide pen coordinate signals and pen up/down signals. These signals are applied to the signal filter and segment integration unit to define segments of strokes that correspond to continuous motion of a pen on a tablet in a fixed direction. This output is next used in the base stroke classification unit which uses a detection process to classify the motion of the pen between pen-down and pen-up occurrences into one of different categories. This classification also indicates whether the stroke has crossed a prior stroke.

- A signal filter and segment integration unit are also part of the on-line recognition system:

  -1. The signals from the tablet are passed to the processor in blocks that correspond to all the strokes of a signal symbol. The fluctuations occur as a result of the tablet's finite spatial resolution and sampling time.

  -2. The filtering algorithm is used to produce new output coordinates whenever the pen position differs from the preceding filter output by a prescribed threshold amount.

- I have studied the basic stroke classification unit and final symbol element recognition unit.

  -The basic stroke and crossing information is analyzed by the symbol element recognition unit which is essentially a basic symbol elements identifier. This operation interprets the basic strokes that have been recorded for the word and then generate a sequence of symbol elements that occur in this symbol.

- The symbol recognition output table shows the sequence of symbol elements that are interpreted in the symbol recognition output table. This table provides a word identification operation in the form of a simple table look-up to determine the word that had been written.

  Figure 2.1 on section 3.1.1 illustrates the above concepts about Chinese Character on-line recognition system between tablet and software.

3.2 Structural method:

Radical extraction methodologies can be classified into two categories: Structural method and Statistical method (such as template matching, partial matching, clustering matching.)[JZ99] I will not go into the details of statistical method in this paper, although statistical method is very useful. This thesis is concerned with the first of these methodologies.

3.2.1 The importance of structural method:

- The structural method provides an intuitive way to decompose Chinese Character into "meaningful" radicals before recognition. Chinese Character are composed of some basic structural components (radicals), Chinese Character radicals, which have special meaning. When Chinese character had been simplified in China by eliminating the radical inside the Chinese Character, the simplified Chinese character survived but lost its meaning and become a strange Chinese character.

- If we can successfully extract radicals embedded in Chinese Character, considerable important information will be obtained. Using this useful information, we can filter out unsuitable radical templates. The number of templates that need to be matched are thereby greatly reduced.

- The complexity of HCCR will also be simplified because recognizing radicals is much easier than recognizing the whole Chinese Character. There are only hundreds of radicals among thousands of Chinese Characters. It can thoroughly overcome the radical translation and scaling problem after normalization.

3.2.2 The detail of structural method

The structural method has the following functions:

- Using the information (length, orientation, and location) of strokes.
- Using the relationships (relative location relation, neighborhood relation, and ordering relation) between strokes.
- Using the relationships (relative location relation, neighborhood relation, and ordering relation) among radicals.
- Using the background information(gap among radicals).

Tree representation of Chinese Character

Both structural method and statistical method have advantages and disadvantages. Ideally, it is wiser to combine the advantages of both two methods. Since Chinese Character are composed of radicals, they can be represented generally by a tree with leaves representing its constituting radicals, intermediate nodes representing its sub-component, and the root as the Chinese Character formed by the radicals in leaves.

For example, the Chinese Character(see Figure 6.1 at the end of Chapter 6)which means "Europe" that consists of five radicals. The reconstruction process is from bottom to top, whereas the radical extraction process is from top to bottom. Now we can propose a recursive hierarchical radical extraction method to resolve the radical extraction problem. This method, which combines the structural and statistical methods, will stimulate the top-to-bottom radical extraction process covering all possible radical combinations.

3.2.3 Concepts and system overview of structural method.

Four main concepts are used in my proposed method as follows:

- Concept 1: Chinese Character structure information

  - For each stroke, the length, orientation, location, number of cross points, number of corner points are extracted;
  - Stroke to stroke relations are the special relations between two strokes and include stroke-to-kanji spatial relations: crossed, connected, left, right, up and down relations.

  - The relative location of each stroke in a Chinese Character.

- Concept 2: Some radicals have stable, salient structural features.

- Concept 3: Gaps exist among radicals. The segmentation is still necessary.

- Concept 4: The cohesion property of strokes in a radical show that strokes belonging to the same radical will be enclosed by a convex hull. Therefore, radicals can be clustered in the center of the associated convex hull.

Structural analysis

Chinese Character radicals can be classified into one of the ten patterns shown as follows:
1.single-element(SE)

2.left-right(LR)

3.up-down(UP)

4.up-left(UL)

5.up-right(UR)

6.left-down(LD)

7.up-left-down(ULD)

8.left-up-right(LUR)

9.left-down- right(LDR)

10.surrounding (SU)

The radicals of patterns in pattern four to pattern ten have stable and salient structural features. These structural features, as mentioned in concept A, are used in order to decompose the Chinese Character into two parts: the expected radical and the remaining components. Once the input Chinese Character is classified as pattern 4-10, the expected radical will be extracted and identified.

Chinese character pattern detection:

In the following context, we will demonstrate the methods of Chinese Character pattern detection used in detail with pattern 8-10

- Pattern 8(LUR): Chinese Character in this pattern have a salient radical located outermostly at the left-up-right area. The detection of this outermost component has to be performed first. Due to the variations of this pattern, it can be divided into four sub-patterns.

- Pattern 9(LDR): In this pattern, Chinese Character have a salient radical located outermost at the left-down-right area. This outermost component is composed of one leftmost vertical stroke, one bottom-most horizontal stroke, and one rightmost vertical stroke. The length of these three strokes are long.

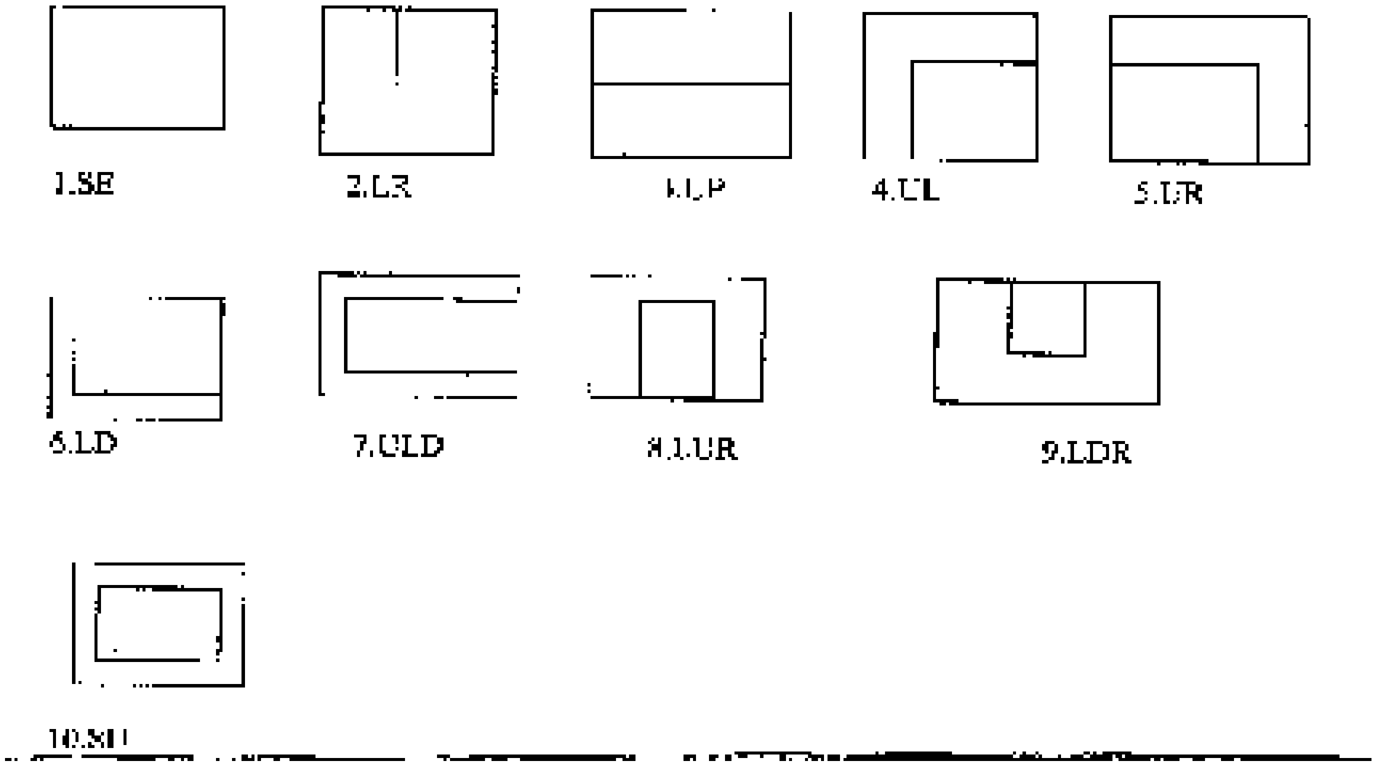


Figure 3.2: Ten Patterns For Chinese Character Radicals[WFW97]

- Pattern 10(SU): The Chinese Character in this pattern have a salient radical located on outermost part at the surrounding area. This outermost surrounding component is composed of one topmost horizontal stroke, one bottom-most horizontal stroke, one leftmost vertical stroke, and one right-most vertical stroke. Furthermore, the lengths of these four strokes are very long.

In the process of Chinese Character pattern detection, some exception to the rules must be invoked to exclude the erroneous cases.

These exceptions rules are problem dependented. For example, the word (see Figure 6.2 at the end of Chapter 6) which means "talkative" with left radical "mouth" and right radical "knife" should be treated differently from an ordinarily word. The stroke s2 (radical knife) does not contain a 4-fork point, so the three stroke s1(radical mouth),s2, and s3 cannot be treated as the component "open mouth"in this module. The right radical "knife" and the left radical "mouth" are the correct radicals extracted by the exception rules.

Straight cut line detection algorithm to achieve the radical separation and extraction goal using straight cut line needs to follow the next steps:

- Step1: The Chinese Character or sub-component will be bounded first by a square.

- Step 2: From the perimeter of the square, trace inwardly the image after the stroke extraction until a pixel is met in some stroke. Then count the tracing depth and record it.

- Step 3: If we can trace from one side to the opposite side, then there exists a clear gap. The wider the gap range, the more stable the gap.

- Step 4: Determine the combination type of gaps after tracing all the gaps.

  1. Type 0- No gap.

   2. Type 1- Horizontal gap exists.

   3. Type 2- Vertical gap exists.

   4. Type 3- Both horizontal and vertical gap exit.

- Step 5: Type 0- go to next layer.

   1. Type 1- input image will be decomposed horizontally into some areas.

   2. Type 2- input image will be decomposed vertically into some areas.

   3. Type 3- input image will be decomposed horizontally or vertically according to the rules.

- Step 6: The areas containing two or three strokes will be checked in advance by radical's knowledge.

  According to some knowledge (number of cross-points and corner-points, the number of horizontal, vertical, slant-to-left, or slant-to-right strokes) about a radical, it can be observed and decided whether it must be merged to the nearest area or remain unchanged. The areas containing one stroke will be merged to the nearest area according to the distance between these two areas.

- Step 7: The resultant areas may contain a radical (or just a sub-component), so we must recursively go back to layer(L) 1 to check whether it needs to be decomposed further. Example word: lake (see Figure 6.3 at the end of Chapter 6) Lines with two arrows (L1 and L3) are straight cut lines. Line with one arrow (L2) is a nearly straight cut line. This word "Lake" is decomposed into four areas by these three vertical cut lines.

Stroke Clustering using K-mean clustering algorithm

The K-mean clustering algorithm tried to resolve the problem in the ambiguous area. Its use in this contex has been proposed by Wang. Fan and Wu.[WFW97]

- Before performing a k-mean clustering algorithm, we must first perform two tasks:

  First strokes connected with four-fork points must be grouped to form stroke blocks in order to avoid ambiguous and erroneous clustering result.

  The 2-D coordinates (xi, yi) of the center points of all stroke blocks in the input component are reduced to the samples with 1-D data (ti) according to the following equations:

  1. For left-right characters or components: $\mathrm{t}_i = x_i, i = 1, 2, ..., n$;

  2. For up-down characters or components: $\mathrm{t}_i = y_i, i = 1, 2, ..., m$;

     where n and m are the numbers of stroke blocks.

- Second, the k-mean clustering algorithm is applied to cluster the reduced data into two clusters by assigning k =2. Example: The circles represent the grouping strokes.(See Figure 6.4 at the end of Chapter 6)

- The redistribution strategies in the ambiguous area of treatment

Although the k-mean clustering algorithm minimizes the sum of square distance from all samples in a cluster to the cluster center, the clustering results may still be erroneous. Therefore, it is necessary to redistribute the misclustered strokes by using redistribution strategies for ambiguous areas which can not be done in this method. This is the disadvantage of structural strict and rigid method.

3.3 Shape classification.

Shape classification is another important concept to know due to disadvantages discovered in structural radical method.[LN93] [XPR00]

- The structural radical method is very 'meaningful' to those who know Chinese Character and write in block style. Otherwise this method will be difficult to apply by the computer. Therefore I propose to move from straight line strokes to shape classification in order to find a more flexible and simplified method without checking Chinese Character's radical.

- As mentioned earlier in this paper, the nature of handwriting presents serious difficulties with respect to personal writing style, structural positioning, etc. Without certain restrictions, it is impossible to achieve machine based recognition tasks. For these reasons, I will base my shape classification on block writing style again. This means that a person is required to write each Chinese Character normally, stroke by stroke, as the beginning students of Chinese character writing do.

- Although the Chinese Character alphabet is very large and each Chinese Character is a complex graphic symbol, the number of different stroke shapes is very limited in the block writing style. This allows us to define some basic shape categories then classify a handprinted stroke into a certain category.

- According to the Chinese dictionary, Chinese Character strokes in block style can de divided into 26 basic categories (1 to 26) and 3 three compound categories (27-29). The compound shapes are usually written by many people as one stroke although they should consist of two strokes.

3.3.1 Locus direction: locus means the exact place of stroke

Locus direction plays an important role in shape classification . Suppose v_I is the current dominant point and v_{I+1} is the next one, we can define eight kinds of locus direction from v_I to v_{I+1}: North, North-East, East, South-East, South, South-West, West, North-West. Angle α is formed by the line segment v_I v_{I+1} and x-axis. It is computed as

$$[\mathbf{tan^1}(\mathbf{y_{I+1}} - \mathbf{y_I}/\mathbf{x_{I+1}} - \mathbf{x_I})]\mathbf{mod\ 360^0}.$$

The central angle is unequally divided into eight parts because the distribution of locus direction of Chinese Character strokes is unequal. For example, it is rare that a Chinese Character stroke is written in the direction from east to west or from south to north in normal written. The following table should be unequally divided into eight parts according the angle range at last paragraph.

Locus direction	**Angle range**
W	170–190
NW	190–260
N	260–280
NE	280–337.5
E	337.5–22.5
SE	22.5–67.5
S	67.5–105
SW	105–170

Table 3.1: Locus Direction Table

This table shows the eight directions and their corresponding angle ranges. The angle alpha is unequally divided into eight parts because the distribution of locus direction of Chinese Character strokes is unequal.

[1] In Taiwan Chiao-Tung University, there exists basic strokes laboratory for public usage once the users or researchers obtained the official permission to use this laboratory.

Chapter 4

Stroke Number free and stroke Order free recognition

The recognition method adopted in this work is Stroke Number free and Stroke Order free. In this chapter I will first describe the reasons for selecting this method. Then I will present the objective to reach and describe this method in detail. Finally I will define the tasks implemented by each module.

4.1 Motivation for a stroke-number-free and stroke-order-free for basic stroke recognition:

Based on the surveyed literature (in excess of 100 research papers), I reached the conclusion that stroke-number-free and stroke-order-free for basic stroke recognition is most suitable for the recognition problem at hand. The reasons to select this method are summarized as follows:

- Chinese Character may be written with large variations in the stroke order, stroke number,as well as shape distortions. They are still readily recognizable by human eyes.

- Most of the recognition methods have constraints on the stroke order or/ and the stroke number imposed on them in order to achieve high recognition accuracy on a real-time basis. It is desirable to develop new methods that can relax or eliminate these constraints.

- In 1983 T. Wakahara and M. Umeda proposed the method that relaxed the constraint on stroke number in their paper titled "Stroke-number and stroke-order free on-line character recognition by selective stroke linkage method"[WakaUmeda83]. The constraint on stroke order was still enforced. They used a dynamic programming (DP) matching method.

- Several other researchers such as K.Odaka, S. Hashimoto, and T. Wakahara proposed algorithms to allow stroke-number and stroke-order variation. T. Wakahara and M. Umeda also developed a Japanese word processor 'AESOP'[WM83] using an on-line handwritten character recognizer (OLCR). This OLCR could recognize Chinese Character for daily usage. It permitted stroke-order variation.

- This method has also been tested in Taiwan where, in 1998, J.W. Chen and S.Y. Lee proposed a US patent using a rule-based approach without having constraints on the stroke number and order. This patent obtained an 80% score rated by the Intellectual Property Network.

- According to my observation, the CHCCR is the same if the tablet and pen are used. Especially for the basic primitive strokes, it depends very much on the classification method. (Just as the graphical primitives depend on different vendor's classification method.) No standard classification could be found yet in CHCCR.

4.2 Objectives to reach in the current implementation:

With the bacground provided by the above observation, the objectives to reach are the following:

- Objective 1: Develop and use a rule-based approach without having constraints on either the stroke number or the stroke order.
- Objective 2: Develop an approach which allows a stroke number or the stroke order free method for on-line basic primitive strokes recognition.
- Objective 3: Use Dynamic programming method to find final solution. The first step is to create several algorithms relying on the Principle of Optimality(P.O.O).

4.3 Summary description of methodology in basic stroke recognition:

The recognition method will make use of several databases as follows:

- **Database of basic strokes**: It will perform a basic stroke recognition procedure to identify all basic strokes contained in the input script.

- **Database of statistical features**:For preliminary classification, we will use database of statistical features. Characters are indexed by stroke number.

- For structural analysis, we need the following five databases:

- **Chinese Character description Database**: This database stores rule codes of constituent components and Chinese Character structures for a plurality of template Chinese Character included in a predetermined vocabulary. Each of such

Chinese Character structures provide spatial relationships between this constituent components of a template Chinese Character.

- **Chinese Character stroke correspondence rules Database**: This database stores strokes Correspondence rules for all the components denoted by the rule codes described in database of character description.

- **Chinese character structures database**: This database stores syntheses rules of Chinese character patterns, decomposition rules of Chinese character structures, and spatial relationships between components for all the Chinese character structures included in the system.

- **Standard component patterns database**: This database stores normalized standard patterns of components.

- **Database of spatial relationships**: This database stores spatial relationships between strokes of each components. An example are the fore strokes of a word. It provides the spatial relationships between fore strokes I and I +1 that described by the following four vectors: $ss(I, I+1), se(I, I+1), es(I, I+1), and\ ee(I, I+1)$.

  These vectors are defined as the connections between the starting point(s) and the end point (e) from stroke I to stroke I+1.

4.4 Detailed explanation about this proposed method

The recognition process can be divided into five steps:

- Step 1: Input of a handwritten script on an on-line basis.

- Step 2: Preprocessing the input script to reduce a number of possible matching template characters.

- Step 3: Performing basic stroke recognition using first of all a database of basic strokes to identify all possible basic strokes contained in the handwritten input script.

- Step 4: Performing stroke correspondence using a database of stroke correspondence rules to find matching strokes in template characters for the strokes contained in the handwritten input script.

- Step 5: Performing computation of discrimination functions using a database of character patterns and a database of spatial relationships between strokes of characters to find one or more template characters with minimum error.

4.4.1 Key features of methodology

To cope with the structural complexity of the Chinese character, strokes are classified iteratively as follows:

- During **the stroke correspondence stage**, strokes are classified into different types of primitives: Fore, back strokes, and points.

  1. **Fore strokes** are those strokes that appear in a character pattern, which can be either a template character or an input script. In the standard templates, the fore strokes should not be connected. In reality, the fore strokes are often connected and the pseudo-segments may appear as fore strokes.

  2. **Back strokes**: The pseudo-segments that connect two contiguous fore strokes are called back strokes, meaning that these strokes should not exit in the standard character. For example, the kanji of common used family name Wang can be easily written with back strokes. Therefore, the back strokes are fictitious strokes which provide connections between fore strokes.

The back strokes allow for the situation that the pen never leaves the paper when it should have. When people write cursive style Chinese character as a result of hasty handwriting, a fore stroke may appear as a fictitiously created back stroke in template. On the other hand, a back stroke, the imaginary stroke connecting two fore strokes, may also be truncated and thus degenerating into an intersection point

3. **Point**: They are identified for the input script as to accommodate connected strokes.

- Each **stroke correspondence rule** contains a specific set of stroke information including:

1. Allowed stroke type.
2. At least one geometric feature measure.
3. Criterion for applying the geometric feature measure: Eight types of strokes correspondences are allowed.

$Fore- > fore, back- > back$

$Back- > fore, back- > point$

$back- > null, null- > back$

$fore < - > null, null < - > fore$

4.4.2 Geometric feature measures

In order to facilitate stroke recognition, a geometrically related characteristic measure based on x or y coordinates, length, or distance is associated with a particular stroke.

Twenty-seven types of geometric feature measures of strokes are defined. They are utilized to establish the stroke correspondence rules. Each rule contains the information that the matching stroke must have a maximum or minimum value of a certain type of geometric feature measures. It is very important to understand the geometric feature measure.

Each geometric feature measure is considered a special geometrically related characteristic length associated with a stroke to aid stroke recognition. Each stroke is considered to be bounded by a Minimum Bounding Rectangle(MBR).

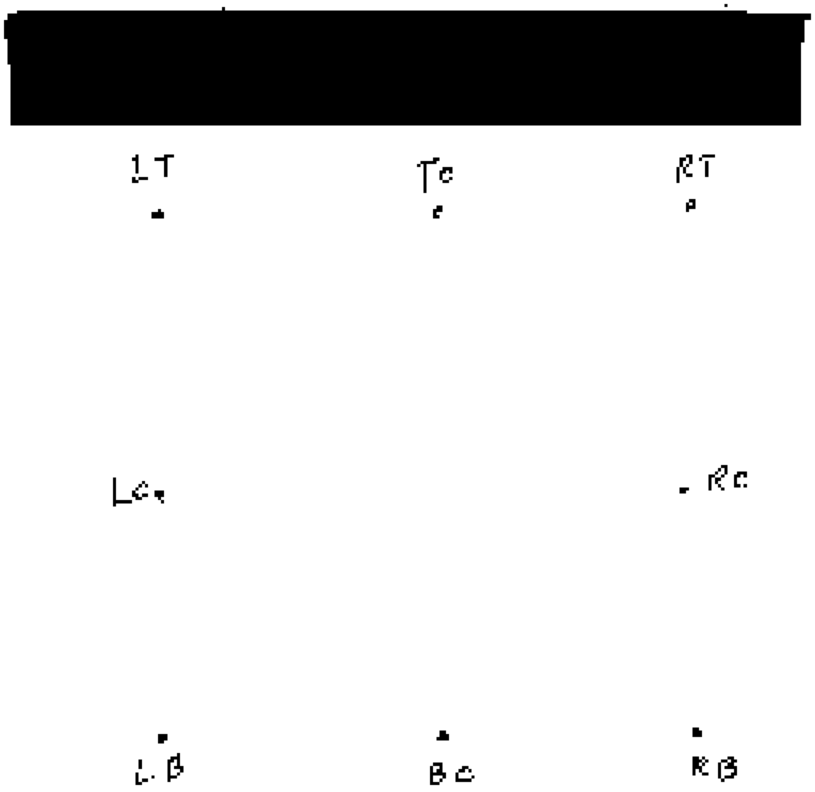


Figure 4.1: The eight Reference Point of the bounding Rectangle of a Chinese Character.B:Bottom C:Center T:Top R:Right L:Left

No.	Geometric feature measure	Remarks
1	MBR_{Xmin}	Coordinates of the four boundaries of a stroke's
2	MBR_{Ymin}	MBR(Minimum Bounding Rectangle)
3	MBR_{Xmax}	and the center point
4	MBR_{Ymax}	of stroke's MBR
5	$MBR_{Xcenter}$	
6	$MBR_{Ycenter}$	
7	S_x	$S = Start\ E = End\ C = Center$
8	Sy	
9	Ex	X and Y coordinates of the start and end points of a stroke
10	Ey	
11	ED (S2, LB)	Euclidean distance from a stroke's
12	ED (S, LT)	start point to one corner point of
13	ED (S,RT)	a decomposed component's MBR.
14	ED (E, LB)	Euclidean distance from a stroke's
15	ED (E, RB)	end point to one corner point of
16	ED (E,RT)	a decomposed component's MBR.
17	ED (C, LB)	Euclidean distance from a stroke's
18	ED (C, RB)	center point MBR to one corner
19	ED (C, RT)	point of a decomposed components's
20	ED (C,LT)	MBR.
21	ED (C, LC)	Euclidean distance from a stroke's
22	ED (C, BC)	center point MBR to one corner point of
23	ED (C, RC)	one boundary of a decomposed
24	ED (C, TC)	component's MBR.
25	L	Length of a stroke.
26	$MBR_Xmin + MBR_Ymin$	block distance measures of the left-bottom corner
26		point of a Chinese character to the left-bottom corner
26		point of an individual stroke's MBR
27	$MBR_Xmax + MBR_Ymax$	block distance measures of the left-bottom corner
27		point of a Chinese character to the right-top corner
27		point of an individual stroke's MBR

Table 4.1: Minimum Bounding Rectangle Table

4.4.3 The Current Method

In my methodology, 14 basic stroke types and 27 geometric feature measures of strokes are used in the design of the stroke correspondence rules.

When the input character script entered is to be recognized, all possible strokes in the input character are identified before the invocation of the stroke correspondence rules.

During the input step, the input script is converted into a collection of sampled points. Each point is described by a vector of pixel(x,y,f) wherein x and y are the coordinates of that point and f is a flag that has a value of on/off. An input stroke is defined to contain sequentially contiguous points with ”on” values. It will end with an ”off” value.

The selected input stroke is matched against the template strokes using the stroke correspondence rule for each template stroke until a matching stroke is found.

The connected strokes are segmented according to their basic stroke types. The geometric feature measures of the strokes become the invariant features of the strokes in 2-dimensional handwritten characters.

Using these invariant features, which are based on the geometric feature measures of the strokes, handwritten words with wide variations in both the stroke number and the stroke order can be recognized.

The computation time required for finding the stroke correspondence between a template character and an input script is the order of O(n) where n is the stroke number of the template.

Each handwritten input script character is first processed during the preprocessing stage, which includes normalization, sampling, and line-segment approximation. This is the main proposed implementation that we are preparing. The input script will match candidate template characters in this stage.

After the preprocessing step, the basic strokes are identified, the estimated range of possible stroke number and the statistical features of the input character are used in preliminary classification to reduce the number of template characters to be processed in the later stage of structural analysis.

The processing steps of structural analysis include stroke correspondence, computation of discrimination functions and detail recognition.

Chinese characters can be handwritten in various styles (font). When the style of character to be recognized changes, I want the correspondence rules in the reference database that are associated with strokes to be changed as well.

The input word is finally identified while using suitable discrimination functions as the template character with the minimum distance. This distance is a measure of **error** between the template character and the input script. When there is more than one template with the same minimum distance, the input character is further identified by traversing a logical tree composed of special structural features.

Standard patterns of Chinese character are constructed based on a number of basic strokes according to certain structural rules. I want to take advantage of these structural rules for machine recognition.

I want my system to contain certain rules, which are predefined in a reference database to effectuate the stroke correspondence (or recognition) between a template character and the input script to avoid the need of repeating combinational exhaustion. For each stroke contained in a template character, one rule is utilized to find its possible matching stroke in the input script.

4.4.4 The step of stroke correspondence

This step performs stroke matching between an input script and a template character. It involves finding a binary relation (r) between the set of template primitives(t) and the set of input primitives(i).

Because the reference database contains only fore strokes, these fore strokes are matched first. As explained previously, a primitive is a basic stroke in a broader sense. It includes the conventional fore stroke and an imaginary back stroke and a "point".

The binary relation r for "stroke matching" from set X to Y is defined as follows: $(r) : X- > Y$

where X denotes the set of primitives of a template character and Y denotes the set of primitives of an input character. For any element $x_i \in X$, if the mapped image (for example a matched primitive) $y_i = r(x_i), wherey_i \in Y$ exists, then there exists only one such image y_i.

If a possible basic stroke in the input script is found to match with one rule, then this selected basic stroke is removed from the input pattern, and the next stroke correspondence rule is applied into the remaining strokes of the input pattern until all the rules in the template are exhausted.

4.4.5 Types of stroke matching

In Figure 3.1, I have already mentioned there are eight types of matching the key features where $x- > y$ indicates a matched pair. A fore stroke primitive is represented as "Fore". "Back" represents a back stroke primitive. A point primitive is represented as "Point";null indicates no matched primitive.

The point primitive only exists in the input script. A back imaginary primitive stroke in the template can be a fore stroke, a back stroke, a point in the input script, or no match. Therefore, the back stroke shows its importance in recognizing Chinese handwritings.

4.4.6 Distance Functions

The distance functions are used for character discrimination defined as the sum of the distances calculated in the various matching types. For example, the function formula $quant8$ is defined to quantize the directions of the vectors into eight direction codes such as $ss(I, I+1), se(I, I+1), es(I, I+1), and\ ee(I, I+1)$.

For the matched pair of$fore- > fore$, the cost function is defined as following:

$$ff\_err\_ss(I) = \begin{cases} 0 & \text{if } quant\_8(ss(I, I+1)) \in ss(I, I+1) \\ 1 & \text{Otherwise} \end{cases}$$

The other cost functions such as $ff\_err\_se(I), ff\_err\_es(I)$ are similarily defined.

Assuming the template character has FM strokes, for the matched pairs of the type $fore- > fore$, the distance function related to the spatial relationship between the fore strokes is defined as:

$$f_{ff}(X, Y) \sum_{i=1}^{FM} = ff\_err\_ss(I) + ff\_err\_se(I) + ff\_err\_es(I) + ff\_err\_ee(I)$$

The matched pairs of the type $back- > point$ occur in the case of connected strokes. The distance function is defined as:

$$f_{bp}(X, Y) = \sum_{i=1}^{BP} 1 (= BP)$$

BP is the number of matched pairs of the type $back- > point$.

The similarity between the template and the input characters is described by the following formula of distance(f$_d$) :

f$_d(X, Y) = f_{ff}(X, Y) + f_{fn}(X, Y) + f_{nf}(X, Y) + f_{bp}(X, Y)\ which\ cumulates\ the\ individual\ results\ o$

The candidate character with the minimum distance in output is the most likely to be the maximum probability of being the correctly recognized character as the input character. If more such candidates exist they are all output as a logical disjunction. (We can also write that the number of the candidate characters constitute a similar group which are represented by a logical tree.)

4.5 One simple algorithm for stroke extraction by finite automata

The below finite automata demonstrates the procedure of processing:

Input: A line segment and the start point for an expected stroke type.

Output: An extracted stroke which is represented as a sequence of line segments if the 2D line segments are accepted by this finite automata.

```
Begin
1. Set current stroke CK to be the line segment with the given starting point;
2. If there is a line segment which is near the starting point of CK and collinear with CK then exit; /*Not starting line segment*/
3. repeat {

    3.1 repeat {
        For each line segment L which is near the ending point of CK
            If L is the legal line segment for this automaton and the angle between L and CK falls under a given threshold, then append L to CK; endif
    } until no other lines can be merged;
```

```
   3.2 go to next state; } until the final state is reached or unable to continue;

   END.
```

4.6 Another simple algorithm for character matching by DFS

The following algorithm describes the major actions taken in the process of Depth First Searching:

- Step 1: Initialization
- Step 2: Recursion
- Step 3: Termination in the final state set.

Input: Extracted strokes in on-line model, stored as a 1D array

Output: The number of matched strokes between the input and the reference template character.

```
Program for matching
Begin
I = 1;/* The pointer to the first stroke type in the on-line model*/

For each candidate stroke J of type on-line(I)
do {
   Repeat{
```

```
 While candidate stroke K of type on-line (I+1) is not the last entry

 Do{

             If (K is dummy) or (the relationship between J and K,

             rel(J,K), is the same as on-line(I+1)

             and the line segments of stroke K do not

             appear in the strokes on the path from the node of stroke K to the       ro

             then


If K is dummy then

  If (miss stroke > threshold T1) then Call backtrack;

  Else  Increase miss stroke by 1;

          Endif

                Endif


     Push J onto the stack S;


             J = K; /* descend to the next level of the search tree*/


             I = I+1; /* Point to the next stroke type*/


             If on-line(I) is the last stroke type in the model

             then Match stroke  = (I+1)/2 /* Searching the missing  stroke

                  Output (Match_stroke); /* The matching succeeds*/

                  Stop;


                  Endif
```

```
            Else /* else of first if*/

            K= next candidate stroke of type on_line (I+1)
            /* The candidate strokes are represented by a linked list*/
           endif /* end of first if*/
      } /* end of while*/

      call backtrack; /* candidate strokes pointed by K are all illegal*/

      } until J is the root;
} /* end for*/

output ('The matching is not successful!');
End.

Procedure backtrack;
Begin

K = J; /* backtrack one node up*/
Pop J from the stack S;
I = I-1; /* Find the preceeding stroke type*/
If the relation between J and K is definite the call backtrack;
Else
K= the next candidate stroke of type On-line(I+1);
End.
```

4.7 Decomposition of radicals

As we know from shape classification correct decomposition of radicals is a difficult but important procedure for an on-line Chinese character recognition without writing constraints (such as stroke order and stroke number). To some extent, the same problem is experienced in the off-line Chinese character recognition system such as OCR system, the extraction of radicals has been investigated by many researchers.

Typically, the background information of a handwritten Chinese character is utilized to separate radicals. But in the conventional approach, radicals that touch each other or are inherently connected cannot be satisfactorily separated. Some researchers, as Cheng and Hsu [CHC89] used the heuristics on the stroke connections, others used relaxation and graph matching methods [CL97] to identify radicals. These methods are all computation intensive.

My method is an improved on-line Chinese character recognition system which includes a hierarchically structures references database which allows the required data storage space to be substantially reduced while providing better Chinese character recognition efficiency. This method also can be applied to generate an infinite number of Chinese character-based print and/or screen fonts.

This improved on-line Chinese character recognition system is a rule-based system. The hierarchical representation of the template Chinese character involves storing the template characters as comprising three major part: Chinese character patterns, spatial relationship between strokes, and stroke correspondence rules, all follow a hierarchical structure.

Chapter 5

Discussion of the contribution of approach and conclusion

Discussion of the contribution:

According to "Ambassador"magazine the May-June/2002 issue(page 9), Chinese Language users make up the highest language population in this world(next is English, Hindi and Spanish). This situation represents a lot of commercial, educational, and cultural needs. With the expectation in future to use the information with CHCCR, it is worth accomplishing this hard research task.

This study is very difficult for the general Western public to understand and it is also very difficult for an ESL(English Secondary Language) student to explain. My contribution of the proposed approach is important for the field. It can be summerized as follows:

- Discovered the Fuzzy Rule-Based with radical extraction method for total general HCCR process. According to researchers Lee, Huang and Shen,[HHS98] this method has an optimistic 99.63 % rate of **divisible Chinese characters** recognition based on 5,401 Chinese character samples. This method did not test undivided Chinese character Recognition.

- Adopted the structural approach for type, position, and radical classification. The structural method provides an intuitive way to decompose Chinese character into meaningful radicals before recognition.
- Noticed the PC and tablet/pen connection problems and using C to resolve this problem.
- Selected the basic and the most important concepts of CHCCR from 100 research papers with all sources correctly cited.
- Discovered the basic strokes laboratory in Taiwan at the University of Chiao-Tung. This niversity has its computer laboratories opened to some Canadian and European Universities that signed a special treaty. The organized research at HCCR in Taiwan and in Japan are very advanced.
- Organized one algorithm for stroke extraction by finite automata and another algorithm for character matching by Depth First Searching.
- Simplified the basic strokes recognition process by order-free and number-free processing methods. Basic strokes should not be the problem in CHCCR. This method takes in consideration a wider style of Chinese character, such as cursive scriptraher than just bold style.
- Discovered the Pampilot Chinese version with HCCR that implemented by two Chinese worker in Japan through IBM and applied for the patent in USA. The rate of recognition is 99.5 % according to their patent introduction papers. The Pampilot's Chinese character recognition system is very simple to use, just as En-

glish recognition commercialized in USA that include inside the system in Pampilot Handheld.

Conclusion:

In the future, natural input may replace the keyboard. This is an ideal way to achieve intelligent man to machine communication. However, the topic of Chinese HCCR is still an open problem due to the difficulties inherent in the recognition of handwritten Chinese character. This is mainly due to the large number of categories, the complexities of the characters, the similarity among different categories, and the wide variability among writers.

During the preparation of this thesis, an intensive study and search was performed. This leads to clearly seeing the HCCR problems using a tablet and pen with the block style writing. These difficulties are different from HCCR using OCR problems. OCR researchers have already accomplished more advanced results, (with the pattern matching study using Directional Element Feature and Asymmetric Mahalanobis Distance can reach 99.42 % rate of recognition [NSS+99]) but only the tablet and pen can really replace the keyboard in the future.

Several approaches and useful techniques have been proposed in HCCR using OCR, such as string matching, dynamic programming, relaxation matching, heuristic matching and combinatorial optimal matching. The progress with the tablet and pen is less advanced than OCR methodologies.

Between four major methods of scripts recognition, the matching technique plays the most important role. Structural radical approach can be very meaningful, but its rigidity makes this approach less useful in reality. Fuzzy Rule-Based with radical extraction methodology play an interesting role in such a situation by offering more flexible range classification and high mathematical level of insight to resolve the problems. Chinese character HCCR needs such methodology to include all the different variations and complicated cases with exceptions. This thesis is limited only to basic strokes recognition for the moment. More exploitation of fuzzy logic and method will be expected for higher level study and testing.

I also find that if people can standardize some primitive stroke types such as in the computer graphic field, and accumulate the progress by sharing the results, the progress in HCCR can be more rapid. Sharing of databases between Taiwan universities led to any type of new method to test of HCCR.

In Japan, the largest public database of handwritten characters is the ETL9B [NSS+99]. There are 2,965 kinds of Chinese character and 71 kinds of Japanese characters. Kana, the first class of Japanese Industrial Standard(JIS) is included. The characters have been written by 4,000 people and scanned as bitmaps. There are 200 samples of each character and 607,200 total character samples are included in the ETL9B. Cooperation between various researchers in the field is needed in order to quickly build a better HCCR system because such inventions will allow for a cheaper price and better efficiency to serve the Chinese community in the world. Although products of HCCR already exist in the world (i.e.,Palmpilot with Chinese character recognition by IBM patent in USA) they remain largely inaccessible due to their cost.

The issues of how to prepare the different testing database and how to simplify the connection and installation process remain subjects for future study. I hope that the use of the tablet and pen will replace the keyboard in the near future, at least on limited applications such as business signature verifications and in post handwritten mail address redistribution service.

Chapter 6

Future Work:Data Type Header File and 4 Chinese Characters

Part A.Data Type header file to test:

We propose the following program in C for the description of different data structures with data types which will be done before basic strokes recognition:

```
typedef struct{ UCHAR Prot;    // counts prototypes in stroke segmentation
UCHAR Lsrt;   // Space insertion in ON or OFF
UCHAR Eras; // Erase with gesture or not
UCHAR Ansr; // Counts cluster buffer entries
UCHAR Buff; // Counts stroke buffer entries
UCHAR Wind; // Counts window buffer enries
UCHAR Pnts; // Counts points in buffer
UCHAR Segm;// Counts segments for the winning scale
UCHAR Segs[SC]; // Counts segments for all scales
UCHAR Edit;// Editing OFF or ON

UCHAR Over;// Overwriting ON or OFF

Char  Bind; // Binding is -1, 0, 1

}COUNT;

typedef struct recpt{

short x;   // x coordinate

      short y;   // y coordinate

      short t; // TRANSITION

}RECPT;

typedef struct segs{

short Dir;   // virtual direction of the segment

short Len;  // Its length in tablet coordinates

short Pos; // Positive displacement

      short Neg; // Negative displacement

}SEGS;

typedef struct coord{

short x;

short y;

}COORD;

typedef struct old{   //contains old points

COORD One[SC];   //coordinates of scale point before current

      COORD Two[SC];   //coordinates of scale point before before current

}OLD;

typedef struct begseg{

char curDir[SC];   // actual direction of the segment for a particular scale

UCHAR fierstP[SC] ; // the point the first for the scale
```

```
SEGS inseg[SC]; // parameters of actual segment
}BEGSEG;
```

B) The pseudocode for PenISR(),PenDown(),PenMoved(),PenUp(), and ExeCom() routine.

PenISR() routine:

```
void far PenISR(void)
{   get new value for pen position and pressure/height;
     determine the state of the pen;
     determine <TRANSITION> variable between 2 consecutive states;
switch(TRANSITION>
{
case AWAY_UP:
break;
      case AWAY_DOWN:
      case UP_DOWN:
       draw the first point of contact;
       AddPoint();
       break;
      Case UP_UP:
       Erase old cursor image;
       Draw new cursor image;
       Break;
      case DOWN_UP:
       Erase old cursor image;
       Draw new cursor image;
       AddPoint();
      break;
      case DOWN_DOWN;

If difference between 2 consecutive positions of the pen is more than one screen pixel then:

      Draw line connecting last and current points;
      AddPoint();
      break;
default:
break;
      }
}
void PenDown(DOWNRET *ret, COORD pen)
{
tick.Segm = 0;
for(scale = 0; scale <SC; scale++)
{ tick.Segm[scale] = 0;
(*ret).oneSeg.firstP[scale] = True;
(*ret).oneSeg.curDir[scale] = 0;
      (*ret).oneSeg.inseg [scale].Dir  = -1;
      (*ret).oneSeg.inseg [scale].Len = 0;
      (*ret).oneSeg.inseg [scale].Pos = 0;
      (*ret).oneSeg.inseg [scale].Neg = 0;
      (*ret).prevPt.One[scale] = Pen;

      (*ret).prevPt.Two[scale] = Pen;

      }

}


void PenMoved(DOWNRET *ret, COORD pen)

{

factor = 1;

for (scale = 0; scale < SC; scale++)

{

de.x =  pen.x - (*ret).prevPt.One[scale].x;

de.y =   pen.y - (*ret).prevPt.One[scale].y;

spacing = (6400/factor/factor);

if((long)de.x * de.x +((long)de.y* de.y) >= spacing)

{Stroke(ret,pen,de,scale);

      (*ret).oneSeg.firstP[scale] = False;

      (*ret).prevPt.Two[scale] = (*ret).prevPt.One[scale];

      (*ret).prevPt.One[scale]  = pen;

      }

      factor *= 2;

      }

}
```

C)More Pseudocode for Stroke(), InserSeg(), LastSeg(),SymRec(), FinProt(),\\
MatchFound() etc should be designed later.\\

For example: Stroke() routine

```
Void Stroke(DOWNRET *ret, COORD pen, COORD de, short scale)
{ if(if vector<de> lies in the right cone)
{
current direction of <de> = RI;
the length of <de> = X coordinate of <de>;
positive displacement of <de> = max (0, y coordinate of, de>);
      negative displacement of <de> = max (0, -y coordinate of, de>)
      }
      else
      if(if vector<de> lies in the lower cone)
{
      current direction of <de> = DO;
the length of <de> = Y coordinate of <de>;
positive displacement of <de> = max (0, X coordinate of, de>);
      negative displacement of <de> = max (0, -X coordinate of, de>)
}
else
      if(if vector<de> lies in the upper cone)
{
      current direction of <de> = UP;
the length of <de> = -Y coordinate of <de>;
positive displacement of <de> = max (0, X coordinate of, de>);
      negative displacement of <de> = max (0, -X coordinate of, de>)
}
```

Part B. Four Chinese Characters for reference purpose at next page.

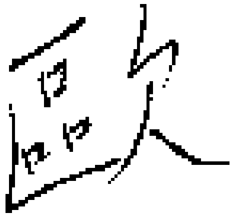


Figure 6.1: Chinese character:Europe

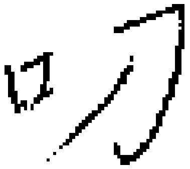


Figure 6.2: Chinese character:Talkative

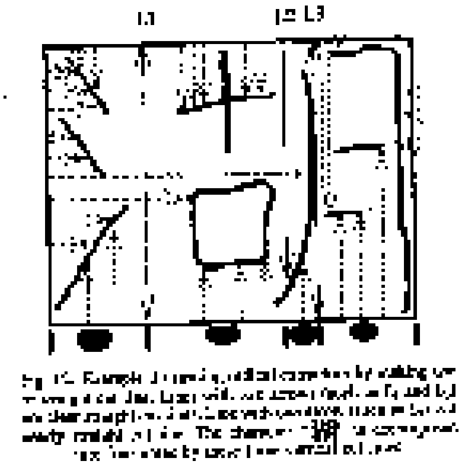


Figure 6.3: Chinese character:Lake[WFW97]

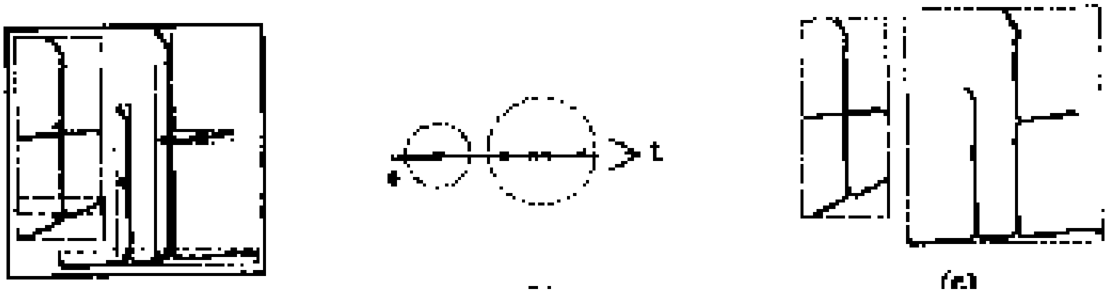


Figure 6.4: Chinese character:Address[WFW97]

Bibliography

[AKF96] A.J.Hsieh, K.C.Fan, and T.I Fan. Handwritten chinese characters recognition by greedy matching with geometric constraint. *Image and Vision Computing*, 14:91–104, 1996.

[A.N01] A.N.Jourjine. Real time handwriting recognition system. *Web Site for Patent Introductory Paper 2001*, 00(0), 2001.

[CHC89] F.H. Cheng, W.H. Hsu, and C.A. Che. Fuzzy approach to solve the recognition problem of handwritten chinese character. *Pattern Recognition*, 22(2):133–141, 1989.

[CL97] C.H.Leung and L.Sze. Feature selection in the recognition of handwritten chinese characters. *Engng. Applic. Artif. Intell*, 10(5):495–502, 1997.

[CLC88] K.J. Chen, K.C. Li, and Y.L. Chang. Stroke-relation coding- a new approach to the recognition of multi-font printed chinese characters. *Computer Processing of Chinese Oriental Languages*, 3:319–330, 1988.

[CT97] H.P. Chiu and D.C. Tseng. Invariant handwritten chinese character recognition using fuzzy min-max neural network. *Pattern Recognition*, 18:481–491, 1997.

[HHS98] H.M.Lee, C.W. Huang, and C.C. Sheu. A fuzzy rule-based system for handwritten chinese characters recognition based on radical extraction. *Fuzzy Sets and Systems*, 100:59–70, 1998.

[JZ99] J.Cai and Z.Q.Liu. Integration of structural and statistical information for unconstrained handwritten numeral recognition. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 21(3), March 1999.

[LJCL90] J.K. Lin, B.S. Jeng, G.H. Chang, and C.J. Lee. Study on the on-line chinese handwritten character recognition. *PostOffice Study Season Review*, 20(1), 1990.

[LN93] X. Li and N.S.Hall. Corner detection and shape classification on-line handprinted kanji strokes. *Pattern Recog.*, 26(9):1315–1334, 1993.

[LW98] C. Lee and B. W. A chinese-character-stroke-extraction algorithm based on contour information. *Pattern Recognition*, 31(6):651–663, 1998.

[NSS+99] N.Kato, M. Suzuki, S.Omachi, H. Aso, and Y. Nemoto. A handwritten character recognition sysytem using directional element feature and asymmetric mahalanobis distance. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 21(3), March 1999.

[RB80] Radio-Beijing. *Apprenons Le Chinois.* Foreign language by Radio-Beijing, 24, rue Bai Wan Zhuang, Beijing, China, 1980.

[SH84] C.Y. Suen and E.M. Huang. *Frequency Distributions of Chinese Radicals and Structural Composition of Chinese Characters.* PhD thesis, 1984.

[SH98] P.N. Suganthan and H.Yan. Recognition of handprinted chinese characters by constrained graph matching. *Image and Vision Computing*, 16:191–201, 1998.

[TSW90] C.C. Tappert, C.Y. Suen, and Toru Wakahara. The state of the art in online handwriting recognition. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, 12(8):1000–1030, August 1990.

[TT98] Y.Y. Tang and Lo-Ting Tu. Offline recognition of chinese handwriting by multifeature and multilevel classification. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(5):556–561, May 1998.

[Wan01] P. Wang. A study of chinese writing systems. From the Series The Artificial Intelligence and Chinese Characters, 2001.

[WFW97] A.B. Wang, K.C. Fan, and W.H. Wu. Recursive hierachical radical extraction for handwritten chinese characters. *Pattern Recognition*, 30(7):1213–1227, 1997.

[WM83] T. Wakahara and M.Umeda. Stroke-number and stroke-order free on-line character recognition by selective stroke linkage method. *ICTP '83*, pages 157–162, 1983.

[XPR00] X.Li, M. Parizeau, and R.Plamondon. Training hidden markov models with multiple observation- a combinational method. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 22(4):371–377, April 2000.

[YF96] D.S. Yeung and H.S. Fon. A fuzzy substroke extractor for handwritten chinese characters. *Pattern Recognition*, 29(12):1963–1980, 1996.