2002

Off-line Thai handwriting recognition in legal amount

Watchara Chatwiriya
West Virginia University

Off-line Thai Handwriting Recognition
In Legal Amount

Watchara Chatwiriya

Dissertation submitted to
College of Engineering and Mineral Resources
at West Virginia University
in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy
in
Computer Engineering

Dr. Ali Feliachi
Dr. Ronald Klein
Dr. Norman Lass
Dr. Roy Nutter
Dr. Powsiri Klinkhachorn, Chair

Lane Department of Computer Science & Electrical Engineering
Morgantown, West Virginia
2002

Keyword: Thai handwriting recognition

ABSTRACT

Off-line Thai Handwriting Recognition

In Legal Amount

by

Watchara Chatwiriya

Thai handwriting in legal amounts is a challenging problem and a new field in the area of handwriting recognition research. The focus of this thesis is to implement Thai handwriting recognition system. A preliminary data set of Thai handwriting in legal amounts is designed. The samples in the data set are characters and words of the Thai legal amounts and a set of legal amounts phrases collected from a number of native Thai volunteers. At the preprocessing and recognition process, techniques are introduced to improve the characters recognition rates. The characters are divided into two smaller subgroups by their writing levels named *body* and *high* groups. The recognition rates of both groups are increased based on their distinguished features. The writing level separation algorithms are implemented using the size and position of characters. Empirical experiments are set to test the best combination of the feature to increase the recognition rates. Traditional recognition systems are modified to give the accumulative top-3 ranked answers to cover the possible character classes. At the postprocessing process level, the lexicon matching algorithms are implemented to match the ranked characters with the legal amount words. These matched words are joined together to form possible choices of amounts. These amounts will have their syntax checked in the last stage. Several syntax violations are caused by consequence faulty character segmentation and recognition resulting from connecting or broken characters. The anomaly in handwriting caused by these characters are mainly detected by their size and shape. During the recovery process, the possible word boundary patterns can be pre-defined and used to segment the hypothesis words. These words are identified by the word recognition and the results are joined with previously matched words to form the full amounts and checked by the syntax rules again. From 154 amounts written by 10 writers, the rejection rate is 14.9 percent with the recovery processes. The recognition rate for the accepted amount is 100 percent.

Acknowledgement

I would like to record my deep gratitude to my advisor, Dr. Powsiri Klinkhachorn for giving me the opportunity to conduct this research and for his invaluable guideance, support, and encouragement throughout. I am also grateful to Dr. Normal Lass, Dr. Roy Nutter, Dr. Ali Faliachi and Dr. Ronald Kline for their undying support, patience and encouragement throughout this demanding project. And a special thank to Cynthia Klekar, for her willingness to assist me in proof reading this manuscript. Without her, this project would still be but a dream.

I am greatly indebted to my parents Wittaya and Malee, my brothers and sister Wiriya, Wisanu and Worawan, and especially to my dear Lina, for their support during my study.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Handwriting recognition is one of the most desirable computer features in enhancing communication between humans and computers. According to initial research, handwriting recognition is the abilities to read and understand human language in written form, and the information is then recorded in digital format for other uses. Handwriting recognition was initially investigated in relation to machine-generated characters with consistent shapes and sizes. Intuitively, a recognition system for those printed characters should be feasible. However, it has been proven that creating this handwriting recognition system is not as simple as it appeared when attempting to give it a level of recognition comparable to that of human ability.

Handwriting recognition is one of the most challenging and oldest problems in computer-related research. Individual handwriting comes in various shapes, styles, and sizes. Additionally, various types of noise and error get included in data acquisition as a result of blemishes on documents. The same written alphabets are different even from the same writer. It is most mysterious how humans can recognize and understand these handwritings, which deviate from the standard model and which they have never seen before. Many hypotheses in computing, linguistics, and biology have been proposed but so far no one can claim a full explanation of how exactly the human handwriting recognition system works. Once people learn how to read, then reading any explicit, good-quality handwriting is easy. After learning from experience, one can read ambiguous or wrongly spelled words and still be able to interpret the meaning of the message. The challenge is how to implement computer systems that can read like humans.

This challenge is not new. Almost at the same time of the invention of the computer in the late fifties, the recognition of printed characters—simplified forms of handwriting—interested the research community (Glauberman 1957; Chow 1957). Many studies have followed since then. It was not until the early sixties that the idea of computers reading handwriting came to attention (Eden 1962). However, it was considered a much more difficult problem. The solution to this problem did not reveal itself easily.

It has taken a long time and much effort to make progress in this area. After forty years of research, the number of research papers related to handwriting recognition has reached into the thousands, and printed character recognition is reaching some mature steps (Nagy 2000). This is reflected in the commercial printed character recognition products currently available in the market. However, handwriting recognition is still an unresolved problem, and no research claims full success in the recognition of general handwriting, especially unconstrained handwriting.

Businesses use filled-forms to facilitate transactions in various ways. Some are filled with handwriting, for example, postal address forms. There are few constraints in writing these forms, thus, the readability of the handwritings are quite varied. The levels of legibility in human handwriting range from easy to read, i.e., well-printed in discrete style, to ambiguous and difficult to interpret without knowledge of context, i.e., distorted in cursive style. To process these documents, humans are required for reading and then entering data or other services for each transaction. This job is tedious to the human operators and prone to error due to the countless transactions each day. Handwriting recognition systems are employed to solve a large part of this problem. It is currently responsible for those documents that are clearly written. The rejected, i.e., unrecognizable units, are bypassed to be processed by humans.

Equipped with powerful computing and fast mechanical systems, the current recognition technology can handle well-written documents in high quantities, and rejecting a manageable amount of unrecognizable units for human evaluators to deal with. Examples of these applications include sorting mail (Cohen et al. 1991; Downton 1992), and reading checks (Suen & Strathy 1998; Gorski 1999; Lecce et al. 2000; Freitas et al. 2000). The studies of fully automatic, human-independent recognition systems are in progress but in reality such systems are only in the early stages. Human reading is still essential in most businesses. Many difficulties regarding handwriting recognition systems so far have no solutions. Some of these problems are even regarded as paradoxical. For example, Sayre's paradox refers to ambiguous character segmentation problems which remain unsolvable to machines, but which can be correctly segmented by average humans. Therefore, the problems in this area remain challenging and solutions will not reveal themselves easily, or at least not in the near future.

The English language uses the Latin alphabet, which is the most studied alphabet in handwriting recognition literature. The Latin alphabet descended from the Etruscan alphabet, which developed from the Greek alphabet. The English alphabet has 52 letters, counting the upper and lower case letters uniquely, 10 Arabic numerals, punctuation marks, and a variety of symbols. The Latin alphabet is one of the most widely used alphabets in the world, utilized in over 30 languages such as Albanian, Dutch, English, French and Vietnamese. Conceptually, the knowledge of Latin alphabet recognition can be directly applied to these languages (with different vocabularies). On the other hand, this knowledge is only partly applicable to other language families, such as Thai, which have different alphabets and writing systems.

In this study, the language of focus is the Thai language. The Thai alphabet is influenced by the Khmer alphabet that descended from the ancient Brahmi alphabet. The other language that has a similar set of alphabet characters as the Thai language, but with a slightly different writing system, is the Laotian language. The Thai language is used by 60 million people in Thailand and the Laotian is used by 25 million people in Laos. In ancient times, these two Southeast Asian countries had strong ties both culturally and politically. In the same way that Latin alphabet recognition technology is shared by all cultures which use the Latin alphabet, Thai recognition research should benefit Laotian recognition technology as well. Currently, knowledge of Thai handwriting recognition is in its beginning stage. However, to facilitate a country's development, it is essential for a country to make progress in its own handwriting recognition technology.

In Thailand, a developing country, it is very common to find computer products—both hardware and software—originating from other nations, such as the US and Japan, that are modified to suit the Thai language and business environment. For most hardware, the modifications seem straightforward, such as voltage or frequency adaptations. In software, localization, such as displaying and printing Thai language, takes more effort and development time. Due to its small market size compared to other more widely used languages, Thai localization of software is initiated and carried on mostly by Thai researchers with little or no support from the software's original producer. Nevertheless, in Thailand software localization is successfully developed in most applications; however, this has not been the case in the handwriting recognition area.

The Thai and Latin alphabets are different in shape, number and set of characters, and in their writing system. Clearly, Thai recognition systems may share similar architectures or processes with Latin recognition systems but the details of its implementation are necessarily different. Some processes need to be totally re-defined and re-evaluated to suit the Thai script. Obviously, it is not possible to slightly modify or localize a Latin alphabet recognition product to be used for Thai handwriting recognition because both languages have entirely different characteristics. A unique Thai handwriting recognition system needs to be designed almost from the ground up, and a deep knowledge of research related to Thai language recognition is required. To accomplish this, not only research in computer recognition itself is needed, but also research in multidisciplinary areas, for example psychology, bio-physiology, and linguistics. Nevertheless, the fact that every essential and fundamental component is currently in its initial state makes a practical, working system still far from fruition. Thai handwriting recognition research has thus far been conducted within very small groups. With a lot of effort, two commercial products with Thai character recognition systems became available on the market in 1985. These products can recognize printed Thai characters with some fixed fonts but their performance lags distantly behind that of Latin alphabet products (Sornlertlamvanish 1999).

The small amount of research reports currently available clearly indicate that Thai handwriting recognition systems are only in the beginning steps and require much more research to make a practical system materialize. There are less than 20 reports published and most of them are conference papers. Moreover, very few of these have been published in international journals over the past thirty years. Most reports are in the area of printed character recognition, with only a few starting to move toward handwriting recognition. In addition, when observing closely the details of each experiment, the measurements of each report are independent, and therefore impractical for comparison with each other. None of the reports declared using some common or standard set of samples. No public database of handwriting exists and no research has been conducted or any investigations made to give guidelines for such a database to be used among research communities. Indeed, the basis for Thai handwriting research is vague and weak. In summary, the existing Thai handwriting recognition research at this point only examines the handling of isolated characters and hand-printed types written by a small group of writers (e.g., not more than 20), and with strict constraints.

Thai businesses employ a numbers of operators to process high quantities of written documents such as banking and postal service forms by reading, entering, and engaging in other services for each transaction. For example, in Thailand 73 million checks were reported in 2000. The approach currently used continually adds more workers to deal with the increasing number of transactions. Intuitively, the increasing amount of transactions will reach a point at which adding workers will no longer be a feasible solution. As Thai businesses continue to grow, this untenable number of written documents will eventually become a problem. The businesses will be forced to look for better solutions to handle the undeniably increasing number of transactions. One of these solutions will depend on an automatic recognition system. Currently, however, they are facing a dilemma as no practical recognition system is working with the Thai alphabet, and handwriting recognition systems are especially lacking. Presently there is neither a Thai handwriting nor a printed character recognition process employed in either the Thai government or Thai private sector industries.

Implementation of a fully automated recognition system in each application is a vastly difficult problem. A system is composed of many complicated components that require an enormous effort to develop. For example, a check reader system to read and record transactions between payee and recipient requires sub-recognition systems for the following units (Tsujimoto & Adada 1992; Gorski 1999). The printed character recognition system is needed for bank name and branch, check number, bank routing number, and payee name. The handwriting recognition system is required for pay date, payee name, amount in number, and amount in word. In addition, the verification systems are needed to verify the payee signature. Because of the difficulties of the recognition problem, studies are usually limited in their scope to a specified area instead of attacking the whole problem at once. In the case of the bank check reader, researchers will narrow their studies to numeric handwriting, word handwriting, printed character, or printed numeric recognition, and limit the vocabulary within the selected scope, e.g., the legal amounts. Then these essential components are combined together to be a complete system.

The aim of this work is focused on finding a recognition model and strategy for unconstrained Thai handwriting. The system scopes are limited to recognizing the amount, or legal amount, used in a bank check when written in words. The handwritten legal amount is required in most monetary documents. Most monetary forms such as bank checks and money order forms often require an amount handwritten in words. The recognition of legal amounts is only one component of a fully automated processing system. However, it is considered one of the most critical and essential tasks, although more complicated and difficult, required in a recognition system. As a brief review of literature given in previous paragraphs reflects, this paper's research can be considered as pioneering a new area in the field of study by specifically focusing on off-line written word recognition in legal amounts. Moreover, the success of this study will have a large impact on improving the quality of services in various public areas. The direct application of the study will be in businesses that use limited vocabulary such as check or other monetary document reading. Indirectly, the benefits of this study should be extendable to other applications that require Thai handwriting recognition for form filled documents as well. In the following section, the handwriting recognition system in general will be discussed, followed by a discussion of the Thai writing and recognition systems. Finally, the proposed research on Thai handwriting recognition of legal amounts will be presented.

1.1 Handwriting recognition systems

Handwriting recognition was developed to allow computers to read and understand human language in writen form. Handwriting is a way to communicate and exchange messages using masks or symbols governed by rules of a language. Language is a way of communicating and exchanging messages. It is defined as "a system of communication consisting of small part and a set of rules which decide the way in which these parts can be combined to produce messages that have meaning" (Cambridge Dictionary Online). In written language, the basic units, that is alphabets and sets of writing systems, produce meaningful messages. Meaningful parts are classified as characters, words, phrases, and sentences (Srihari 1990).

Writing environments include writing equipment and constraints. The writing equipment is a marker, such as a pen or pencil, and the written object or media, such as paper or leather. Examples of writing constraints include time, writing space, angles of writing, etc. These environments directly affect the appeal and readability of the writing. The basic rule is the more constraints there are, the easier the reading. In addition, writing for oneself has fewer constraints than writing for others, therefore, ones own writing read by someone else.

Individual handwriting is not consistent and has no definable writing rules, thus, the recognition system must be able to handle such variations and generalizing natures. Each individual's writing deviates from the standard or ideal model of characters and inconsistency. The difference comes from many factors such as individual practice and biophysical reasons, i.e., brain, hand, and muscles structures (P1amondon 1990, 1995). According to John Favata, "Some studies suggest that handwriting styles are not even stable over relatively short intervals of time (page to page) for an individual author" (1993). Both machine-generated type and human-generated writing recognition systems commonly deal with the similar nature of problems in writing systems. The boundary between these recognition systems is due to the flourish and inconsistency in human writing styles.

There are two categories of recognition systems: on-line and off-line approaches. The on-line approach will sample pen movements in real time and record the positions and activities (e.g. pen-up and down) in a timely order. The recording device is often called the tablet digitizer. The off-line approach works with written documents. This approach will digitize the image with an optical sensitivity device such as a scanner or digital camera. The on-line approaches are expected to register the recognized units immediately or shortly after data acquisition, while off-line approaches might work with a document previously digitized and stored in batch processing fashion. It is generally assumed that an off-line approach is more difficult because in this approach the image of the whole page of a written document is presented for analyzing without knowledge of the sequence of drawing strokes (Plamondon & Srihari 2000; Madhvanath & Govindaraju 2001). The recognition applications for large quantities of transactions have to deal with already written documents, i.e. check filled forms. Thus, they are working in the off-line approach. The problems of the off-line approach are the interest of this study.

Handwriting recognition systems research is one branch of general recognition research. Similar to general recognition systems, handwriting recognition systems have three sub-processes: pre-processing, recognition, and post-processing (see 0). The pre-processing includes noise-elimination, binary-image transformation, text location, line separation, and other processing that is deemed necessary to improve the recognition. General discussion on these topics will be given in chapter 2. The second step is the recognition process, which includes the segmentation, features extraction, and the matching process. The designing of the recognition process typically involves defining a set of features and matching these features to the model to sort out the most compatible candidate. The class models are built by automated training or other methods. In the third step, the outputs from this recognition process are then checked by contextual postprocessing for highly unlikely decisions.

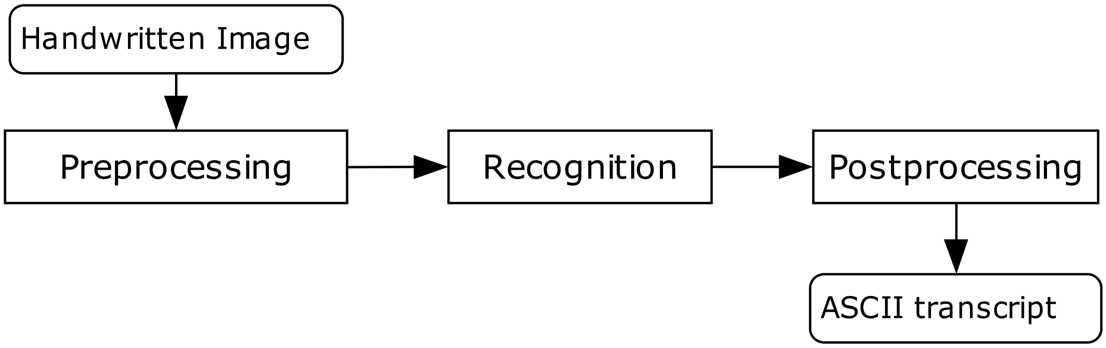


Figure 1.1 General architecture of handwriting recognition system.

Three major components that influence the complexity of a handwriting recognition system are the readability of writing style, the number of writers, and the size of the vocabulary used (Guillevic 1995; Vinciarelli 2000). The handwriting recognition system must overcome these different types of problems. In practice, to solve such a complex problem, some constraints are put on the problems. With more constraints on these components, better overall recognition-rates can be attained.

Although the level of readability of a written document cannot be formally defined, the readability level can be ranked from easy to read by average people, such as hand-printed, to some that are only readable by the writer or cannot be read without contextual information. Shuniji (1998) describes the highest difficulty level in character recognition as that of unconstrained writing having various sizes, shapes, and distortions in critical areas of characters. The handwriting recognition systems, which handle hard to read documents, are intuitively more complicated. In addition, it is very common to find that sometimes people cannot read some of their own writing, for example, writing done in a hurry, or the writing of a short note. These situations indicate the difficulty inherent with the problem.

The variation of styles depends on the number of writers. The larger the number of writers, the more fluctuations in their writing styles. This can bring about more difficulty for the recognition systems. It is a necessary goal that systems for public services are able to read and understand unconstrained handwriting from multiple writers. The ability of handling such various types of input is known as system generality. The system generality can be tested and evaluated practically by sampling a large number of handwritings taken from multiple writers.

The size of the possible vocabulary encountered has direct effects on a recognition system's performance. The large number of possible vocabulary in general that appears in unspecified documents (wherein there are no constraints on the vocabulary allowed, e.g., lecture notes and newspaper) causes the recognition system great difficulty because the vocabulary is virtually unlimited. In addition, confusion measurement among the strings inside the lexicon is the other important factor affecting system complexity (Mohumed 1995). For instance, the similarity between "and" and "ant" often results in misinterpretation.

1.2 Approaches in handwriting recognition

The lexicon-based strategy uses the limited contextual information in a system to reduce one of the problem's complexities. There are two approaches dealing with different sizes of lexicons in handwriting recognition (Shihari 1990; Madhvanath & Govindaraju 2001). The first approach is called the analytical. This approach regards a word as a collection of characters. The word image is segmented into character units and then these units are identified. In the final stage, the system vocabulary, i.e., lexicon, will be used to determine the complete word. The second approach, often called the holistic approach, uses the whole word image as input. The handwriting recognition engine uses the whole word's features to determine the word's identification from the lexicon.

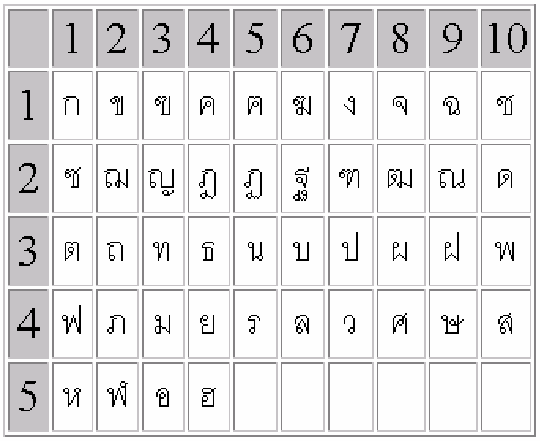
Because of its adaptability, the analytical approach is applied to systems which have medium (between 100 and 1000) and large (more than 1000) lexicons. The recognition engine always deals with the same character classes regardless of the size of the lexicon. An example of the analytical approach is reading city and street names in a postal context, which are medium lexicons. An advantage of this approach is that it only requires training (building the system model) for the characters, not the entire lexicon. However, this approach is based on the character segmentation process, which is difficult and imperfect (Vinciarelli 2000). As stated in Sayre's Paradox (1973), "a letter cannot be segmented (correctly) before having been recognized and cannot be recognized before having been segmented." Many cases of recognition failure originate from faulty segmentation (Casey & Lecolinet 1996).

The holistic approach is motivated by the psychological studies of human reading, which indicate that humans use features of word shape such as length, ascenders, and descenders in reading (Madhvanath & Govindaraju 2001). This approach is applicable for a small lexicon system such as legal amounts or some specified applications. It takes the whole word as input, without a character separation process. This approach bypasses the deficiency of character segmentation problems and shows greater computation efficiency than previous approaches in small lexicon applications. Moreover, it has successful readings in many cases where analytical approaches have failed (Madhvanath 1997). The disadvantage is obvious. It is impractical to build numerous models for entire word classes for unspecified documents. In addition, it should be noted here that the inputs for this approach must be isolated words, as it is only capable of recognizing one word at a time. The existing word segmentation algorithms for English are acceptable with some constrained writing, as English sentences are written with isolated words (Seni & Cohen 1993; Mahadevan & Nagabushnam 1995). However, these segmentation algorithms are not applicable to languages written without boundaries between words in sentences.

1.3 Thai writing and recognition systems

Ancient Thai is in the Tai language family. It was invented in 1283 C.E. Thai script has evolved through many generations since its inception (Comrie 1990). The current script is descended from the King Narai script, which has been used for more than two hundred years. Thai script has a large set of characters comprised of 44 consonants, 18 vowels, 4 tones, 2 diacritics, 3 punctuation symbols, 4 symbols for ancient words, and 10 numerals. Table 1.1 shows 44 Thai consonants. Thai characters can be divided into groups of similarly formed figures (Ronnakiat 1997). Many non-Thai have some difficulty in distinguishing the characters in the same groups, even if in printed form (Doug 1997).

Like Latin-based languages, Thai is alphabetic, written from the left to the right, from the top to the bottom, and in block style. In other words, Thai script is non-cursive and there are no connections between letters. When comparing Thai script with that of the Latin family languages, many obvious differences exist. Thai script has two major unique characteristics. There are no gaps between words in the same sentence or phrase and, secondly, there are four writing levels in one line. These characteristics imply that a great deal of research specified for the Thai recognition system needs to be done before a practical system can materialize.



	1	2	3	4	5	6	7	8	9	10
1	ก	ข	ฃ	ค	ฅ	ฆ	ง	จ	ฉ	ช
2	ซ	ฌ	ญ	ฎ	ฏ	ฐ	ฑ	ฒ	ณ	ด
3	ต	ถ	ท	ธ	น	บ	ป	ผ	ฝ	พ
4	ฟ	ภ	ม	ย	ร	ล	ว	ศ	ษ	ส
5	ห	ฬ	อ	ฮ						

Table 1.1 Thai character sets, only the consonants are displayed here.

In 1983, the first research study examining printed Thai character recognition was published (Kimpan et al. 1983). Since then, activities involving Thai recognition research have continued. However, research made towards Thai handwriting recognition is small in number and superficial. Following the initial study, 3 papers were published in 1993, 5 papers in 1995, and 3 more in 1998. Moreover, most of the literature concerns printed or hand-printed character recognition.

These research documents propose a number of features and recognition models used for Thai handwriting and evaluate the system performance in terms of the correct recognition rate. However, it should be noted here that it is not sensible to numerically compare these recognition results. These experiments were conducted based on different sets of data collections with a different number of writers (mostly with small groups or an unspecified number of writers) and different constraints (mostly with strict constraints). Also, the distribution of the character classes was mostly omitted from the tests and therefore was useless for comparison. There are no claims of using common data sets in any reported experiments so far.

In order to make comparisons between each approach and avoid self-bias, Thai handwriting samplings should be collected, analyzed, and then utilized to create a public database designed as a standard training or testing set. A good database should be created from samples of a large number of writers with various styles to represent the real situation in which the application will be used. Unfortunately, such a database does not exist. It is one goal of this study to create a database of Thai handwriting samples of legal amounts that represents writer-independent handwriting.

In regards to a practical system, Thai handwriting recognition systems are only in a beginning state and far from being realized. Thai printed character recognition, on the other hand, has shown some progress. Two commercial products in Thai printed character recognition became available on the market in 1998 (Sornlertlamvanich 1999). These products are designed for Thai printed characters with specific fonts and sizes.

However, their performances distantly lag behind that of Latin alphabet printed recognition systems. Moreover, there are no reports of any Thai recognition applications being used for public services such as postal or banking services.

Besides the complexities of the problem, the few research efforts addressing this topic are partially due to the lack of local business incentives, the small scope of economically viable applications for the research, and the lack of a supporting infrastructure necessary to conduct and develop a working system. Thai writing and reviews of recognition systems will be further discussed in Chapter 3.

1.4 Problem statements

The difficulties in Thai handwriting recognition mainly arise from two aspects. First, as is common among every written language, there is a great variety in individual styles of handwriting. Second, the recognition systems are complicated by the unique characteristics of Thai script. There is no direct method to compare the difficulty of the recognition of characters between languages, especially in handwriting. However, based on printed characters, some studies show that Thai characters have even less distinguishable characteristics than that of English characters (Doug 1995). This demonstrates why research in this area is difficult.

All handwriting recognition is extremely challenging. The deviation from the model of characters set in sizes and shapes is the primary complication. In addition, some types of writing styles, such as the cursive styles in English, contribute more complexity to the system. There are no specified types of handwriting styles in Thai. That is, Thai handwriting does not divide itself into upper and lower case writing or distinguish between cursive or hand-printed writing as does English handwriting. Thai handwriting is a more relaxed writing system compared to Latin writing systems. This creates more difficult problem to the Thai handwriting recognition systems as well: a structural features distinguishing one character from another might be omitted. For example, a head, the curl at the beginning of a written character, is often omitted in many cases. Secondly, crossing and overlapping strokes between successive characters is common. These relaxed forms make Thai handwriting look like cursive scripts. Some varieties of Thai handwriting are shown in Figure 1.3.

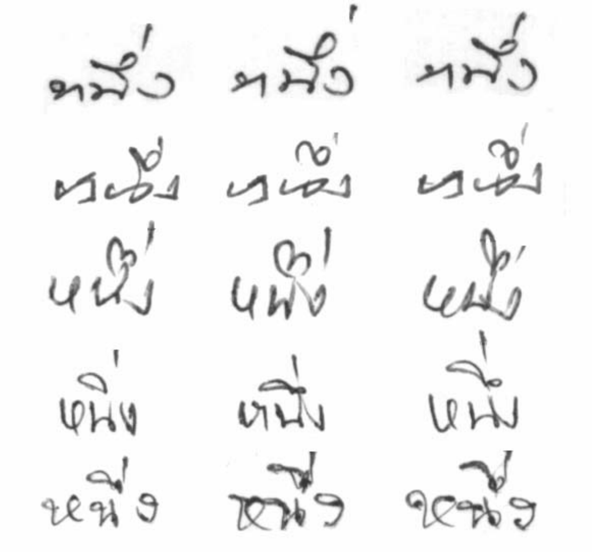


Figure 1.2 Thai handwriting samples from five volunteer.

Other problems arise from Thai writing uniqueness. The necessary assumption for most Latin alphabet handwriting recognition systems is that isolated recognizable units are given. These units are isolated as characters in the analytical approach or as words in the holistic approach. For the Latin alphabet system, there are many acceptable methodologies for handwriting recognition systems to determine the boundaries of characters (Casey & Lecolinet 1996) and words (Seni & Cohen 1993; Mahadevan & Nagabushnam 1995). In regards to segmentation, it should be noted that character segmentation algorithms for cursive English are only partially applicable to the nature of Thai character writing, and English word segmentation algorithms are not applicable to Thai writing at all.

The character segmentation algorithms for Thai scripts are based on the assumption that the input is unconnected written unit (Airphaiboon 1996; Lohakan 1999), but that is only true for hand-printed writing. Generally, this approach will fail in handwriting where connected, overlapped and touching strokes are often found. With respect to word separation, the assumption that the gap between words is actually greater than that in-between characters in the same word resulted in an applicable algorithm for English (Seni & Cohen 1993; Mahadevan & Nagabushnam 1995). Thai writing simply has no gap between words. Therefore, it is very difficult or impossible to separate words from a sentence as shown in Figure 1.3.

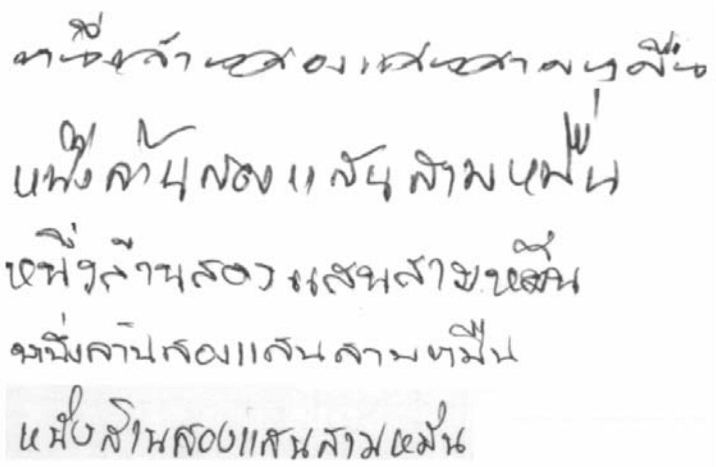


Figure 1.3 Thai handwriting of **หนึ่ง-ล้าน-สอง-แสน-สาม-หมื่น** (1,230,000).

Moreover, as mentioned before, some strokes are actually found connecting the preceeding letters in Thai writing. These touching strokes add more difficulty to word segmentation analysis as well because it can make the whole sentence or phrase appear as one connected unit. In Thai text processing, the word segmentations from a sentence are mostly done with lexicon or contextual knowledge.

To the best of the author's knowledge, there is no research literature published that is applicable to an algorithm or that proposes a strategy for Thai word segmentation in handwriting. Lacking applicable character and word segmentation to deal with Thai sentences or phrases, the existing application of both analytical and holistic recognition systems based on isolated recognizable units will definitely fail. It is one of the goals of this study to find a new model or strategy to be used with Thai handwriting recognition.

The objective of this research is to find a new model and to implement algorithms and processes needed for recognition of off-line Thai handwriting for legal amounts from general writers. The handwriting is unconstrained in size or angle of writing but required to be written in a specified space. The following topics are within the interest of the research: the relevant features to identify Thai handwriting word, the recognition models and processes for the specified interests, and the necessary postprocessing for the specific area of legal amount recognition. The proposed model is based on developing and generating the reasonable amount of hypothesis from available prior knowledge and evaluating by applying each hypothesis to the problem.

The basic goals of this research are:

A) Design and collection of Thai handwriting samples in legal amount. In order to develop a general handwriting recognition system, a large amount of quality training and testing of sampled data from a large number of handwriting is necessary.

B) Analyze Thai legal amount in regard to vocabulary, syntax, and character classes to be used in designing the recognition system. The syntax rules and vocabulary analysis will be used in hypothesis word generation and in spell check in the post-processing stage.

C) Design and implementation of Thai handwriting recognition that uses an image of single character as input. This step will attempt to evaluate a set of features and recognition methods from previous studies.

D) Design and implementation of Thai handwriting recognition that uses an image of single amount word as input. This study will attempt to evaluate the potential holistic approaches in recognizing the handwriting word.

E) Design and implementation of word hypothesis generation module using the prior knowledge from the vocabulary, the available recognition results of character recognition, and the syntax analysis. This study will attempt to implement a system that can generate a list of sensible candidate words and word boundary positions. The basic assumption using this concept is that enough recognizable characters are available. However, this module must be robust enough to handle difficult cases when fewer or very small recognizable characters are presented.

F) Design and implementation of post-processing modules. This module will be responsible for correcting spelling errors. The syntax will be checked if the results seem correct. A reject will result if the wrong spelling or wrong syntax is more than the designed threshold.

G) Design and implementation of Thai handwriting in legal amount recognition system. By combining all expected modules as explained previously, the recognition system for Thai handwriting recognition using the proposed model will train and evaluate by testing a set from the data set.

1.5 The proposed model

The motivations to propose a model to solve the problems described in the previous paragraphs stem from psychological evidence and the potential advantage of word recognition models. Psychological studies state that when reading, words themselves are the readable units, especially with cursive writing: "words were found to be identified under conditions in which their component letters could not be identified. These findings have led several researchers to suggest that readers may in fact sample the visual data in reading rather than process every element of text" (Smith 1969 cited in Guillevic 1995). These studies focused on the English language; however, it is highly likely that this concept applies to other character-based languages such as Thai.

Another psychological study in regards to human reading suggests a different interesting concept. Even that studies, however, is not committed to an exact model of the system. There are indications that the natural reading process is not static but an interactive system able to select a meaningful recognizable unit. McClelland and Rumelhart, in their 1981 study, state, "Humans do not follow monotonic paradigms, feed forward recognition system. They can alter their course of action dynamically and select the utility relative to their goals. The word perception model in the cognitive approach is assumed an interactive process" (Park 1999). This concept implies that instead of a top-down process, a hierarchical and recursively deterministic process could be considered to solve the recognition problem.

The second motivation came from the advantage of the holistic recognition approach. The holistic recognition approach is applicable for handwriting where character segmentation is ambiguous and each individual character is distorted, but the shape of the word can be used as a guideline for identification. The holistic approach is helpful when individual characters cannot be identified but the overall shape of the word is maintained (Madhvanath 1997). Algorithms based on the holistic paradigm are computationally more efficient as well (Madhvanath & Govindarju 2001).

A simple algorithm for separating units within a Thai sentence or phrase image is developed based on "contour following" (Lohakan 1999). These units may be isolated or connected components composed of characters, sub-characters, words, and sub-words. Only the units that are isolated characters or sub-characters conceptually can be identified by character-based recognition, while the units that come from other connected parts remain unidentified. However, when the previous separation results, there is no information indicating the beginning and the ending of the words.

The holistic approach can apply only when the beginning and the ending of words are identified. This is where the hypothesis and evaluation can take part in problem solving (Lin 1980, Win 1984 cited in Favata 1993). This occurs when the hypothesis and evaluation combine with another constraint, such as the limited size of the lexicon or system syntax, and then it is possible to generate a list of candidate words from previously recognized characters and a list of hypothesis-parts of word images as well. In the holistic recognition approach, this information can be evaluated for the best candidate.

From this evidence, and based on an assumption that there are some characters that can be successfully isolated and identified, a hierarchical architecture of character and word recognition levels is proposed to solve the Thai handwriting recognition problems. The flow in each level of this model is similar to that of the classical recognition system. However, the hierarchical system uses both the character and word recognition engines interchangeably. The proposed model is shown in Figure 1.4.

The first step in the process is segmentation and recognition at character level, assuming that some characters can be successfully segmented and recognized. The design and development of the Thai handwriting character recognition system is the first part of the problem. The characters recognized are joined together to form the full amount by matching with the words list or lexicon. The amounts created are checked as to if they are a fit along the character results. The complete amounts are sent to check if their syntax is correct.

If the character segmentation and character recognition processes work perfectly the lexicons matching process should reveal the correct word and should pass the syntax checking. However, if the segmentation cannot separate the connecting character and the character recognition cannot identify those components correctly, the word matching could fail and/or not pass the syntax check. The word recognition system is designed to solve this problem by generating some hypothesis of the word boundaries and segmenting that image. These hypotheses word images are identified by the word recognition. This concept is based on the assumption that even if some characters are connected or broken, the main shape of the word still retains enough information to be identified. The results of word recognition are sent to create the amounts again. These amounts are checked as to if they conflict with the syntax in the last stage.

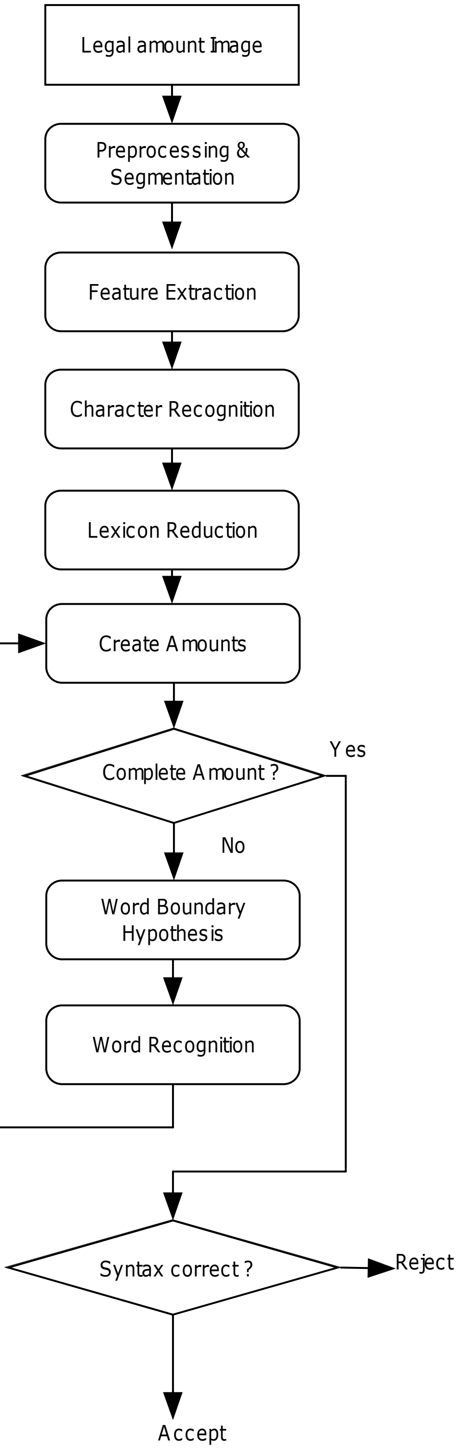


Figure 1.4 The proposed model for Thai handwriting recognition in legal amount.

1.6 Research contribution

This research will contribute to the development and analysis of methods and demonstration of their performance for the recognition model for Thai handwriting in legal amount. This is the first time that a generic database for Thai handwriting recognition in legal amount words will be constructed. The research will be conducted in a manner to realize the multiple writers of legal amounts, which could be extended by adding a large number of writers to represent a writer-independent model. This is the pioneer study to work with a writing system that does not contain boundaries between words.

The expected contributions of this work can be summarized as follows:

- The design and implementation of a data set for Thai handwriting legal amounts.
- A new hypothesis and evaluation scheme for Thai word segmentation for small lexicon.
- The design of Thai handwriting character and word recognition when the isolated words and characters are the input.
- A new recognition architecture which hierarchically combines character and word recognition level that is expected to overcome the problem of incomplete character and word segmentation.

Chapter 2

Overview of Handwriting Recognition Systems

In this Chapter, the literature of handwriting recognition systems is investigated. The scope of the literature study will be the off-line handwriting recognition system for the English language using the Latin alphabet. This review will give the basic idea of how a general off-line recognition system for handwriting works. The focus will be the various types of feature extraction processes and recognition processes. Due to the differences between the Latin and Thai alphabet and their writing system, the recognition systems of these languages are also different. However, the Thai handwriting recognition system can be seen as an extension of the recognition of the Latin alphabet system. The extension of the Thai handwriting recognition system will be discussed in Chapter 3.

Handwriting recognition transforms the artificial marks of languages into symbolic or code representations. Handwriting is converted to digital form by two approaches: the on-line and the off-line approach. The off-line approach uses a light-sensitivity device, e.g., a scanner or a digital camera, to read a written document. The result is color or gray level intensity of the writing document. The second approach, on-line, uses a tracking device to collect time-position-action of writing strokes, i.e., tablet digitizer or pen device. The sequences of the writing positions are stored in a timely order. The off-line approach converts the document into the image for the recognition system. The image shows pixels belonging to writing strokes with no temporal information available. There is no information indicating the beginning or ending of stroke writing. Though the immediate recognition results are always preferred, it is not required in many applications. The data acquisition process can prepare and store the images separately while waiting for the recognition process. The off-line data acquisition and recognition approach are the interest in this study.

The on-line handwriting recognition systems use sequences of a writing stroke. Using a given sampling period, the on-line approach derives higher features, such as directional, velocity or accelerates. Due to the nature of timely stored data, the sequences of the writing are clearly represented, from the beginning through the end of the writing. The word "on-line recognition" also implies that the recognition results are expected shortly or immediately after the writing. The on-line approach has a niche application in personal use devices, e.g., PDA (Personal Digital Assistant). Other related applications of on-line handwriting recognition include gesture recognition and signature verification, etc. A survey by Plamondo and Srihari (2000) gives a summary on the progress in this application area.

2.1 Off-line Handwriting recognition process

The off-line handwriting recognition systems have as input the image of a handwritten document and produce as output the recognized units, i.e., characters or words. The operations in handwriting recognition systems are based on those of classical recognition systems that have many sequential sub-processes. Common processes or modules of classical recognition systems are pre-processing, recognition and post-processing. The pre-processing stages process are the image processing process to convert the input image into preferred format and improve the image quality, for example, convert gray or color image into binary image and removing noise generated from the scanning process or noise from the background texture. The feature extraction process extracts the relevant information, known as a feature vector, which could be used to identify the input image in the recognition step. The recognition process uses these features to find the most compatible class with the input. The postprocessing uses priori knowledge, i.e., component decomposition rules, language grammars, or lexicon, to evaluate and correct the recognition results. Figure 2.1 shows the diagram of a handwriting character recognition processes.

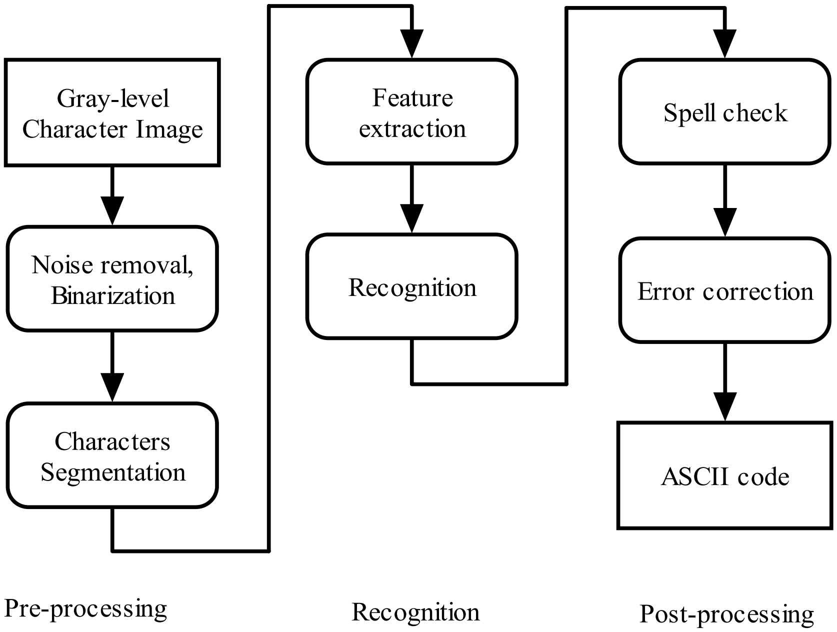


Figure 2.1 Off-line handwriting recognition system processes.

2.2 Pre-processing

The aim of the pre-processing step is to improve the overall quality of the input image by removing the irrelevant elements of the image while preserving the relevant information. This process applies to the whole page of an input image. Common operations of the preprocessing stage include the task of converting a grayscale or color image into a binary image and noise removal, the extraction of the foreground texture and removing noises. The noise removal processes employ methods from the image processing research area.

The grayscale or color images are needed to transform into binary images. This is because a full range of grayscale or color is not necessary for many applications in documentation analysis. This process is often called binarization or thresholding. The basic idea is to assign binary values according to a threshold value. If a pixel has a value greater than this threshold value, it will be assigned to black (1), or else it is assigned to white (0). The task of assigning a threshold is to determine an optimal value to be used.

There are global and adaptive threshold techniques. Global threshold will choose one threshold that applies to the whole image. Adaptive threshold will use different thresholds for different areas by computing a local threshold for each sub region. Researchers from the University of Oslo and Michigan State University compare and evaluate published adaptive thresholding methods on hydrographic charts. The best result on hydrographic charts comes from Niblack's method, based on a threshold set below the mean gray level of a 15x15 window by a fixed fraction (0.2) of the standard deviation of the gray level (Trier et al 1997; Trier & Jain 1995; Trier & Taxt 1995; Taxt et al 1989). Liu & Srihari (1997) proposed a threshold selection for textured background for postal address readers. The process consists of 1) preliminary binarization based on multi-modal mixture distribution, 2) texture analysis with run-length histograms, and 3) selection of the threshold using decision tree.

2.3 Segmentation

Various types of components, such as texts, pictures and graphical marks, comprise a document image. The segmentation process at high level separates the text image from the document or page image. At the low level, the segmentation process separates the recognizing units, e.g., characters or words, from the text image for further processing. At the higher level, region segmentation separates text area from a page layout.

The region segmentation of these components can be categorized into page layout analysis and form processing. Page layout analysis has little or no prior knowledge of the layout and content of the page before processing the analysis. The form processing will be given the reference position of the area of interest. For details of form processing, see Yu & Jain (1996). An example of this information is the position of the writing cells in a page in reference to the edge of the form. The text images can be extracted precisely from the cell while the rest of the irrelevant information can be discarded. This process is known as a form dropout. Line segmentation of a small range of skewed images can be accomplished by examining the horizontal projections profile. However, the actual handwriting might go up and down which makes the line more difficult to segment. Kim et al (1999) use the concept of imaginary-based lines, which form the boundaries that the words of the line need to stay inside of.

English words are separated by space. The words boundaries are found by observing the valleys in the vertical projection profiles. This concept can be applied to printed character separation as well. However, character segmentation for the handwriting cursive scripts is more difficult. The cursive writing style always connects each character in a word with links, which confuses the separation algorithms. Many algorithms are proposed for the problem but so far none claim full success in character segmentation.

Casey & Lecolinet (1996) surveyed several approaches that had been proposed since 1959 for segmenting touching or fragmented character strokes. The degree of difficulty depends on many factors such as the typeface and print-source, as well as on the ratio of font size to scanner resolution. Casey and Lecolinet also defined dissection as the attempt to divide the image into classifiable units. See Casey (1996) and Dunn (1992) for good surveys of character segmentation techniques.

2.4 Feature Extraction

Feature extraction is the heart of the character recognition system. Devijver & Kittler (1982) define feature extraction as the problem of "extracting from the raw data the information which is most relevant for classification purposes, in the sense of minimizing the within-class pattern variability while enhancing the between-class pattern variability." Ideally, extracted features should provide uniquely relevant identification information of character class without redundancies. Features can be extracted from sub-letters, letters or words. These categories are called low, medium and high level features accordingly. Tier et al (1996) review feature extraction methods for character recognition. Feature extraction methods can be grouped into reconstructive and non-reconstructive features. Techniques discussed here are mostly for a single character.

2.4.1 Reconstructive features

Features that can be used to construct the original image are called reconstructive features. The transformation and series expansion feature can generally construct the original image from the feature set. These features are obtained from coefficients of various orthogonal decomposition methods by the representation properties of the image data. Geometric transform, Zernike transform, Fourier descriptor and Wavelet descriptor are the examples of reconstructive feature extraction methods.

2.4.1.1 Geometrical moments

Hu (1962) introduced geometrical moment invariants, which do not vary in translation, scaling and rotation, and are widely used in many applications such as aircraft identification (Dudani et al (1997). Kim &Yuan (1994) have applied these moments to the task of rotation of character recognition. Reiss (1991) revised some of the theoretical proofs in Hu (1962). Hu's geometrical moment invariant features can construct from the central moments. However, the Hu's moment invariants have some drawbacks. One drawback is the increase in complexity with the increase order. The second drawback is that they are not derived from a family of orthogonal functions, and so contain much redundant information about an object's shape.

2.4.1.2 Zernike moments

Teague (1980) proposed the Zernike moments based on the theory of orthogonal polynomial to overcome the problems associated with the regular moments. Teh & Chin (1988) evaluated the Zernike moments among Hu's moments, Legendre moments, Fourier-Mellin moments and pseudo-Zernike moments. The result shows that Zernike has the best performance. The orthogonal moment features have been used for the recognition of Arabic numerals. These are complex valued moments, which are defined using Zernike polynomials. The more terms using Zernike's moments, the more precise the reconstructed image to the original.

2.4.1.3 Fourier Descriptor

The Fourier transform takes a signal in time domain and represents it as its series expansion coefficients in frequency domain. Fourier transform has the ability to produce a signal transformation into an orthogonal space, but the signal must be periodic. To overcome this problem, Zayed (1993) proposed Discrete Windowed Fourier Transform (DWFT) to calculate FFT over a window of discrete size at different temporal points in the signal. Elms (1994) applies the Fourier series expansion to the recognition of printed characters using Hidden Markov Model (HMM) with a single pixel-wide sliding window. The resulting 64-term spectrum is reduced to the first 32-dimensional feature vector.

Zahn & Roskies (1972) use the cumulative angular function (CAF) created by one revolution of tracing to define a windowed signal, which is expanded using Fourier series expansion. Granlund (1972) proposed the contour as a periodic function by calculating the function.

$$b[k] = \frac{1}{L}\sum_{m=1}^{L} y[m] e^{-jk(\frac{2\pi}{L})m} \tag{2.1}$$

$$a[k] = \frac{1}{L}\sum_{m=1}^{L} x[m] e^{-jk(\frac{2\pi}{L})m} \tag{2.2}$$

Where L is a measurement of the distance traveled around the contour. If the length of the entire contour is L, then the function is periodic with period L; k is from 0 to L-1. The rotation and translation invariant can be computed from 2.3.

$$r(k) = \sqrt{|a[k]|^2 + |b[k]|^2} \tag{2.3}$$

Where k is 1 to L-1. For the application of scale invariant, Fourier descriptor can be calculated from 2.4.

$$S[k] = \frac{r[k]}{r[1]} \tag{2.4}$$

Fourier descriptor s[k] are symmetrical around position (L-1/2) and S[1]. The higher order of coefficients represents the detail of the image structure, while the lower order represents the main or fundamental image structure. For a simple structure image, 5 to 8 Fourier descriptor coefficients bring a good approximation of the original image. Figure 2.2 shows the original image of "3" on the first row and the reconstructed image from 9 to 3 Fourier descriptor coefficients on the second row. Figure 2.3 shows the original image of "C" in the first row and the reconstructed images from 9 to 3 Fourier descriptor coefficients on the second row.

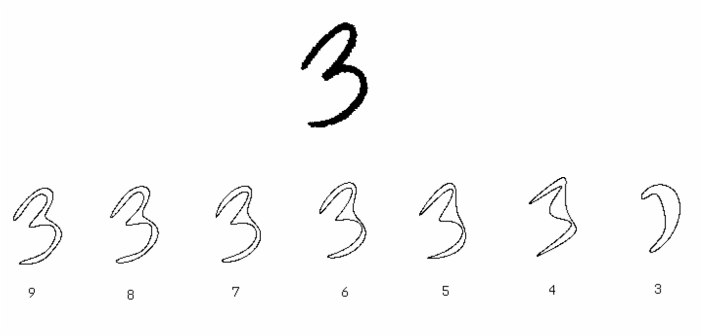


Figure 2.2 Original image of "3" and reconstructed images.

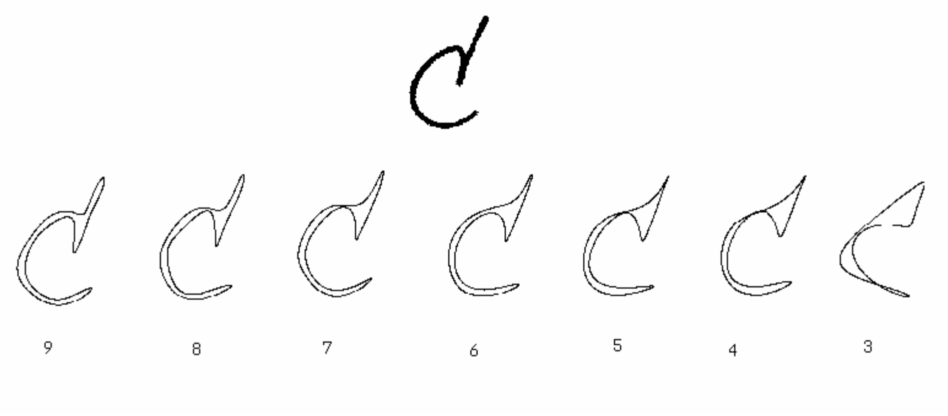


Figure 2.3 Original image of character C and reconstruct images.

Wang et al (1994) proposed the moment Fourier descriptor. This technique calculates the Fourier coefficients for a number of line segments found by searching in a number of directions from the centroid of an object in the image. These coefficients have the ability to represent an arbitrarily complex object while remaining invariant to translation, rotation, and scaling.

2.4.1.4 Character Image

Character image itself can be regarded as the "feature vector". This approach is one of the earliest techniques in pattern recognition research. Assume that the character image and templates share the same reference frame. The dissimilarity measurement (D) between the character image (Z) and each template (Tj) are commonly computed by a mean-square distant (Equation 20.1-1 in Pratt 1991):

$$D_j = \sum_{i=}^{M} \left(Z(x_i, y_i) - T_j(x_i, y_i)\right)^2 \qquad (2.5)$$

This method is simple; however, it has an obvious disadvantage. This method is not applicable to models with various types and styles, especially when taking into consideration handwriting image, which has significant shape, size and rotation variants. However, the character images contain redundant information, so the character's skeleton is widely used instead. An image processing called thinning process is used to reduce an image to its skeleton. Burr (1981) and Wakahara (1993, 1994) use the character skeleton as a feature. In their approach, each template is deformed in a number of small steps, called local affine pattern as shown in Figure 2.4. The number and types of transformations before a match is obtained can be used as a dissimilarity measure between each template and the input pattern. The problem with this technique is that it is not clear how the initial positions of the template were chosen. Moreover, if all possible positions in the image were to be tried, then the computation time would be enormous.

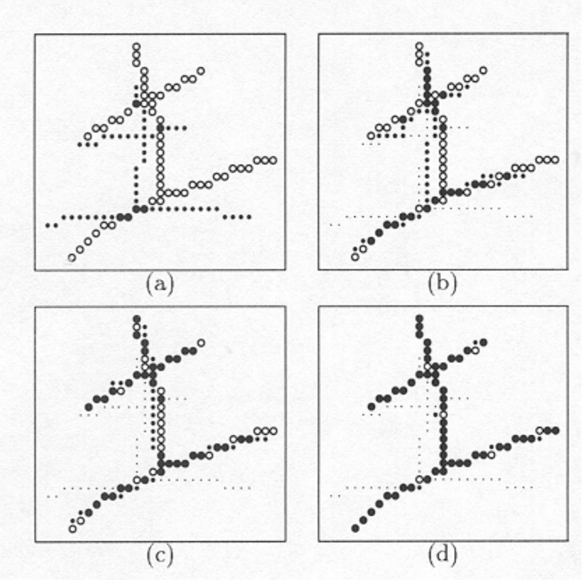


Figure 2.4 Reformation steps to recognize a Japanese character, reproduced from Wahara (1994).

2.4.2 Non-reconstructive features

Non-reconstructive features are mostly the features that can be extracted from the structural or topology of the handwriting globally or locally. These features are expected to distinguish class of character or word. This method is simple in computation and has a reasonable discrimination power. However, these features cannot construct the perfect original image. Structural features had been widely used by many applications (Takahashi 1991; Mohiuddin & Mao 1994; Kundu et al 1989). The characteristics of each character are observed and algorithms are developed to detect those identifiable structures such as loop, ascender, descender, curvature, end line, T and X-join, crossing, and projection profile.

Most of the non-reconstructive features approaches suffer from image transformation, i.e., scaling, rotation or translation. The normalization step is needed to reduce the effect of these transformations. The objective of normalization is to reduce the variance in writing factor of the same characters or words, while maintaining the relevant features of the class of character or word as much as possible. Some recognition models group the normalization processes in the pre-processing step. Basic types of writing variance independent of class identity include height, width, and slant of the writing. Image scaling can perform the height and width normalization to the reference frame. Some techniques (Yanikoglu and Sandon 1994) use different scales for different zones of a character.

2.4.2.1 Projection profiles

Glauberman (1956) introduced projection profiles into the hardware Optical Character Reader (OCR) systems. In additional to recognition tasks, it has been used extensively in printed character, word and line segmentation. The basic projection profiles used in recognition are horizontal and vertical projection profiles. For a horizontal projection, the histogram is the counting of the number of pixels that have corresponding values (0 or 1) along the horizontal axis. Similarly, for the vertical projection, the histogram is created from the counting of the pixels along the vertical axis.

Figure 2.5 shows the projection profiles of number 5. The horizontal projections are shown under the image and the vertical projections are shown on the right. Projection profiles are sensitive to image transformation. It requires normalization processes to adjust the image into a reference frame. The printed character recognition matches projection profiles of a character with profiles of known classes to find the most compatible class. This feature can be used for handwriting recognition systems as well.

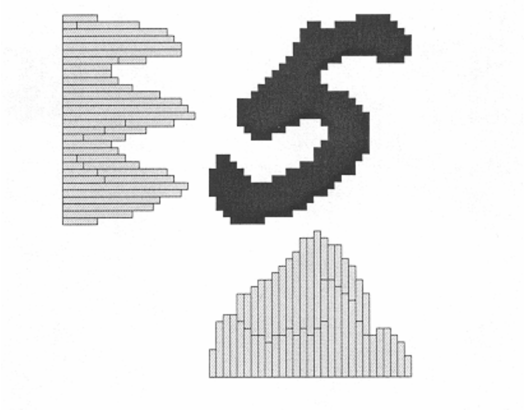
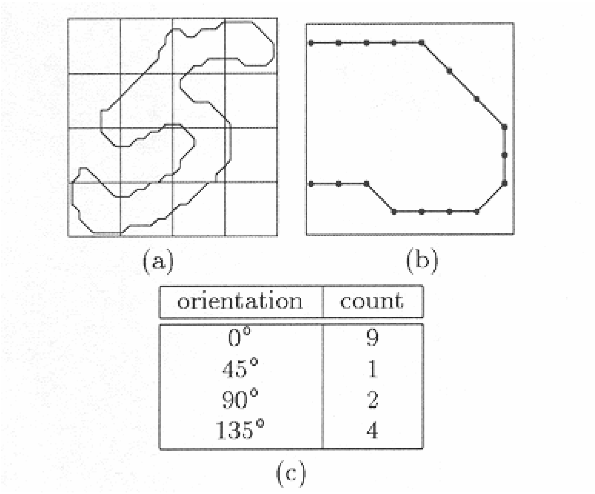


Figure 2.5 Projection profile of number 5, reproduced from Tier et al (1996).

2.4.2.2 Structural features

From a human viewpoint, each character has its own unique structural identification. It is not difficult for a human to identify some type of structure that exists in a character or a group of characters and to conclude the character's identification. This approach is called structural detection. An unknown input is searched for the existence of features -both the quantity and measures of the quality- to identify to class of the input.

Examples of structural detection used in recognition are the number of pixels in sub-area or zone; the number or the existence of the end point, cross point, and t-point for skeleton image; and the existence of various types of cavities in sub-area or zone. Figure 2.6 shows how to apply zoning to a character. The character is zoned into 4x4 sub-regions, and in each sub-region some features of interest, such as orientation of links between contour points of the images, are counted.



orientation	count
0°	9
45°	1
90°	2
135°	4

Figure 2.6 Zoning of contour curve. (a) 4x4 grid superimposed on a character. (b) Close-up of the upper right corner zone. (c) Histogram of orientations for this zone, reproduced from Tier et al (1996).

2.4.3 The recognition process

The recognition process uses the feature set of the handwriting image as input to determine which model class has the best similarity for the input. The four best-known approaches for pattern recognition are template matching, statistical, syntactic and neural network. Template matching is comprised of measuring the similarity between input image, typically a 2-dimension shape, and a group of prototypes or templates. Statistical classification uses a decision function based on the feature set of input image and the representation of feature space of different classes. Syntactical recognition views a picture as a language description and a class as sentences that belong to the language. A specific class can be derived according to a grammar. Neural network is a massively parallel computing scheme having some organized structure to learn non-linear input and output relationship.

2.4.3.1 Template matching

Template matching is the simplest and easiest approach in pattern recognition. It is used to determine the similarity between two entities of the same type. In template matching, a template, prototype or model of the pattern to be recognized must be available. The input pattern is matched against the stored templates with the respect to transformation variations. The result is the measurement of similarity or correlation. The general or global templates may be created from the training set. Template matching is effective in some applications but requires extensive computation power. It also has a number of disadvantages. For example, it has a small toleration for the distorted or view pointed changes, or large various types of intra-class. Deformable template (Burr 1981) or elastic template (Wakahara 1994) matching can be used to match patterns when deformation cannot be explained or modeled directly.

2.4.3.2 Syntactic approach

The syntactic pattern recognition approach regards a pattern as the construction of simple components following rules or grammars. The simple components are themselves constructed from simpler sub-components. The lowest component level is called a primitive. The highest level pattern is represented in terms of the interrelationship between these primitives (Fu 1982 ; Pavlidis 1977). In syntactic pattern recognition, a structure of patterns is analogous to the syntaxes of a language. If the primitive is represented by characters, then the pattern can be viewed as a word or sentence composed of the character following the grammar or rules. This approach is used in situations where the patterns have a definite structure that can be defined in terms of a set of rules, such as EKG waveforms, texture images, and shape analysis of contours (Fu 1982). However, the disadvantage of this method includes the difficulty of segmenting noisy patterns to detect the primitive, the construction of the grammar from training data, the requirement of a vast amount of training data, and an enormous computation time to train the system (Perlovsky 1998). Further study of the syntactic approach can be found in Fu (1982) and Pavlidis (1977).

2.4.3.3 Statistical approach

With the statistical approach, each pattern is represented in terms of *d* features or measurements and is viewed as a point in a d dimensional space. The decision making of which class the input should belong to is made from the boundary in the *d*-dimension features space where *d* is the number of selected features. This depends, however, on whether the *d* feature can create disjoint regions for each class in the feature space. In the statistical decision theory approach, the decision boundaries are determined by the probability distributions of the patterns belonging to each class (Duda & Hart 1973). See details of statistical classification reviews by Jain et al (1999).

2.4.3.4 Neural Network

Neural network is a network of computing structures to learn or adapt according to a relationship between input and output. The common neural networks for classification are back-propagation. For character recognition, the image of the whole character is used as the input to the network during the training step, and its class is presented as the target. Then the network will adjust the network's parameters to reduce the difference between the actual network output and the target value. The training step will stop when the difference is less than some threshold value. Then the network is ready to classify the unknown input.

The most commonly used family of neural network for classification tasks are feed-forward network, the multilayer perceptron and Radial-Basis Function (RBF) networks (Jain et al 1996). For feature selection or class cluster analysis, self–organization map (SOM) networks by Kohonen (1983) are always used. Figure 2.7 shows a character recognition system using the neural network approach. The input to the neural network is a gray level image of a character and the output of the system is the class of numerals from 0 to 9. The neural network gains popularity from its low dependence on domain specific knowledge and due to the availability of the learning algorithms for practitioners to use. Lately, the neural network in recognition is regarded as a non-linear discrimination function for an input-output relationship. On the surface, neural network appears different from the statistical approach. But most of the well-known neural networks are implicitly equivalent or similar to classical statistical pattern recognition methods. Table 2.1 shows compatible functions between the statistical and neural network approaches.

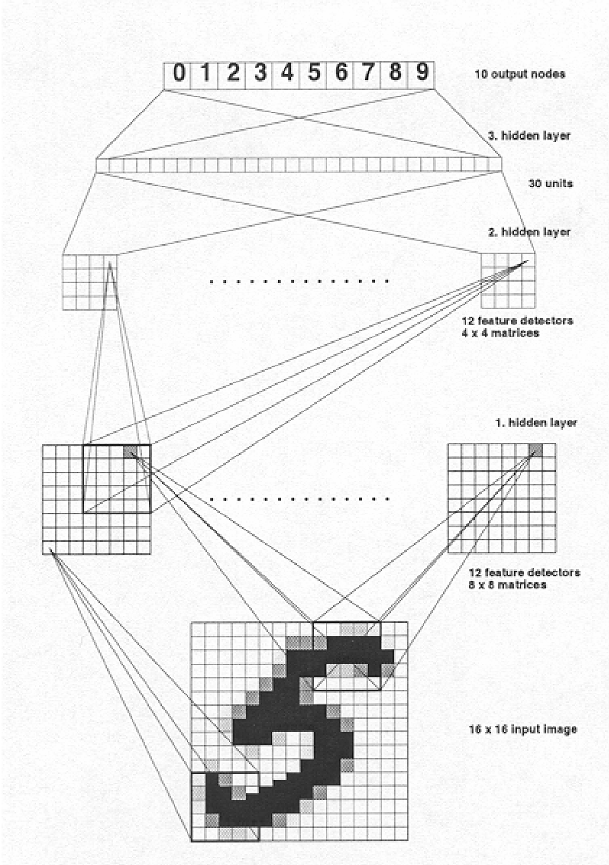


Figure 2.7 Neural network for handwritten zip code recognition, reproduced from Cun et al (1989).

Statistical Approach	Neural Networks Approach
Linear Discriminant Function	Perception
Principal Component Analysis	Auto-Associative Network
A Posteriori Probability Estimation	Multilayer Perceptron
Non-linear Discriminant Analysis	Multilayer Perceptron
Parzen Window Density-based Classifier	Radial Basis Function Network
Edited K-NN Rule	Kohonen LVQ

Table 2.1 The relationship between statistical and neural network approaches, reproduced from Jain et al (2000).

2.5 Post-processing

Results of recognition process are characters or words obtained from the image of segmented recognizable units. In many cases, it is obvious that the recognition of a segmented unit will definitely fail because of an incorrect segmented result or distorted writing. Moreover, there are many cases in handwriting when the same symbol is used to represent more than one character when used with a different word. Human reading of handwriting heavily relies on contextual knowledge (Mohamed 1996). The recognition in character or word level can be improved using other knowledge than from the document image itself. Post-processing uses priori knowledge at a higher level than information from document image, i.e., lexicon, component decomposition rules, or language grammars, to improve the integrity of recognized characters or words.

If the system's lexicon is given and other major characters in a recognizing word can be recognized correctly, then the misrecognized character can be replaced with spell check algorithm. Spell check is most useful with a system where context knowledge is provided. This can significantly improve the recognition rate beyond the information given by a single character. Component composition in specific languages is an essential part of a recognition system (Faure 1996). Many languages have multiple part alphabets, e.g., the letter "i" which has two parts to prevent false recognition: the body and the head. Most of the recognition systems recognize one segmented component at a time, so the recognition result will be the head and the body of letters "i". Post-processing is responsible for concluding the recognition sequence of the head and body components to identify as one letter "i". The system grammars will be very useful when system has limited syntaxes, such as word amount or legal word recognition. The syntax provides a list of possible candidates of unknown words and characters, which is very useful in improving the overall recognition rate.

2.6 Summary

In this chapter a review of off-line handwriting recognition systems for the Latin alphabet is given. Two parts of the handwriting recognition system are emphasized, the feature extraction process and the recognition process. The feature extraction process can be separated into constructive and non-constructive feature types. The constructive features can be used to compose the original character, while the non-constructive features cannot. The recognition process can be separated into template matching, syntactic, statistic and neural network approaches. Direct template matching can be applied to print handwriting but not to the running styles. The syntactic approach offers an advantage when handling new variations of the writing scripts or un-seen writing scripts, but it suffers from the exponentially growing writing rules in real situation. The statistical approach seems to be the more promising approach when dealing with unconstrained handwriting systems. The statistical approach uses many statistical analysis methods to classify data but it requires specific knowledge. The neural network methods, which are the popular methods for data analysis, are in fact the tools that simplify the syntactic approach and allow the researcher to deal with the problems instead of the analysis techniques.

Chapter 3

Research in Thai Handwriting Recognition

In order to understand the nature of the problems in this study, an introduction to Thai language, including an explanation of its character set and writing system, is first discussed, followed by an outline of previous research of Thai machine-generated writing recognition and of Thai human-generated writing recognition. The review of Thai recognition systems will emphasize the separation of characters and the features used in the recognition process.

3.1 Introduction to Thai language

The ancient Thai language is in the Tai Language family, which includes languages spoken in Assam, northern Myanmar, Thailand, Laos, northern Vietnam and the Chinese provinces of Yunnan, Guizhou and Guangxi. Thai script is one of ancient languages, invented in 1283 by King Ramkhamhaeng of Sukhothai. It is called the Sukhothai Script. The Sukhothai scripts are shown in Figure 3.1.

The root of Sukhothai script comes from the Bhram alphabet Grantha, a script that originated in South India (Hudak) and which is also influenced by the Khmer alphabet. In 1357, a new script called the Li Tai came to be used in ancient Thailand. The shapes of the letters in the Li Tai script, the second-generation script, evolved from the Sukhothai script, although some of them were modified. In 1680, during the reign of King Narai, the third generation of the script, called Narai, was brought into use as shown in Figure 3.2. Initially, in the Sukhothai script, Thai writing had only two levels in a line. Then, during the Ayutthaya dynasty, more writing levels were added. The number of the characters and the character shapes of the Sukhothai script are different from those of the present Thai script. However, Thai natives can read and understand most of the 200-year-old documents written in the Narai script without difficulty. For a detailed history of the Thai writing system, see Ronnakiat (1997). The Narai script continues to be developed and preserved as the national Thai script in the present. Figure 3.3 shows handwriting using Narai script.

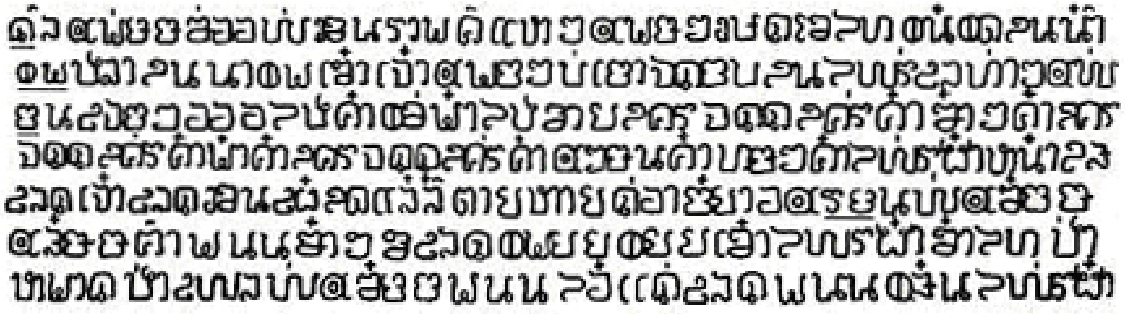


Figure 3.1 Sukhothai script.

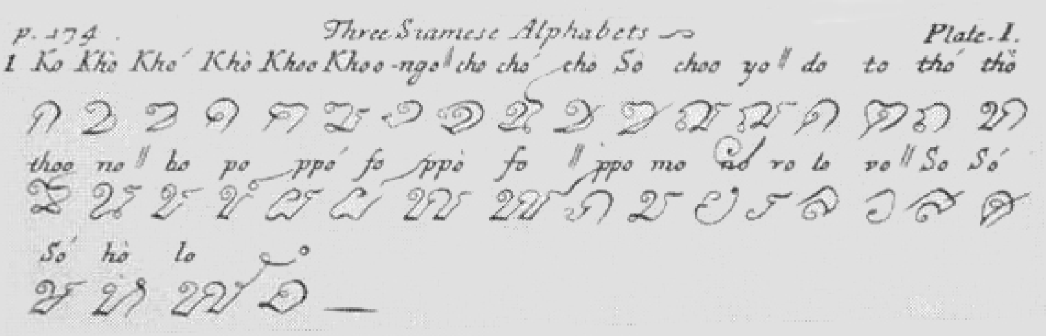

p. 174 .
Three Siamese Alphabets
Plate-1.
1 Ko Khò Khó Khò Khoo Khoo -ngo cho chó chò So choo yo do to thó thö
thoo no bo po ppó fo ppò fo ppo mo ÿo ro lo vo So Só
Só hò lo

Figure 3.2 Ancient Thai consonants: “Nari” script in 1686, reproduced from Navikamoon (1993).

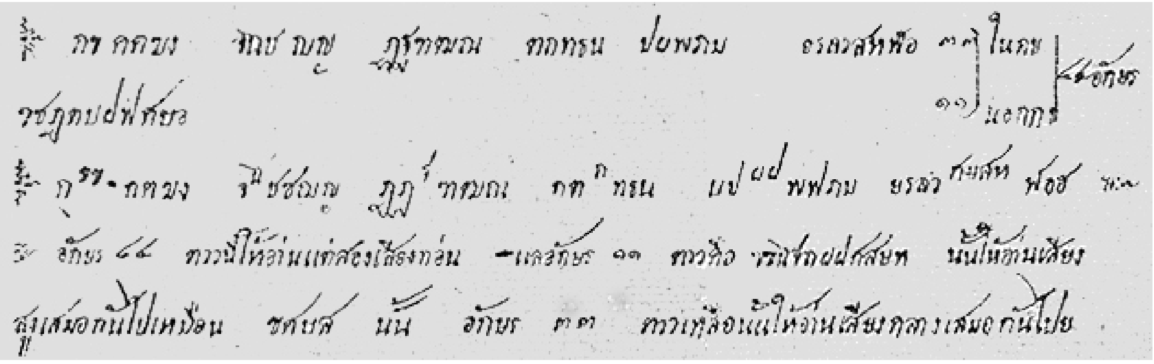
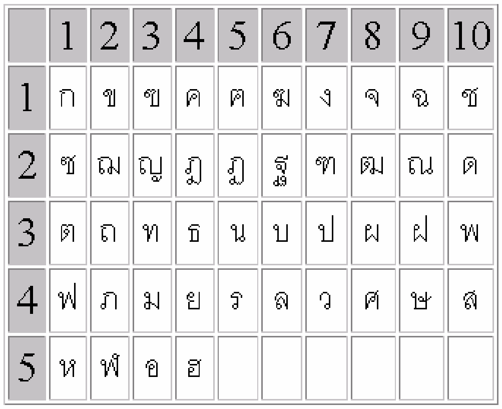


Figure 3.3 Thai handwriting in 1732, reproduced from Navikamoon (1993).

3.2 Thai alphabet set

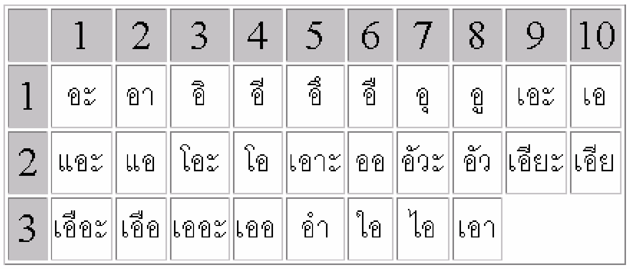
The Thai alphabets consist of 44 consonants, 18 vowels, 4 tone marks, 2 diacritics, 3 special symbols, 4 symbols for ancient Pali/Sanskrit words, and 10 numerals. Consonants are shown in Figure 3.4. There are two obsolete consonants, [ฃ] and [ฅ]. In Thai writing system, there are no capital letters, no punctuation marks and no spaces between words. However, a break may be introduced between groups of words at the writer's discretion (Campbell & Shaweevong 1957).



	1	2	3	4	5	6	7	8	9	10
1	ก	ข	ฃ	ค	ฅ	ฆ	ง	จ	ฉ	ช
2	ซ	ฌ	ญ	ฎ	ฏ	ฐ	ฑ	ฒ	ณ	ด
3	ต	ถ	ท	ธ	น	บ	ป	ผ	ฝ	พ
4	ฟ	ภ	ม	ย	ร	ล	ว	ศ	ษ	ส
5	ห	ฬ	อ	ฮ						

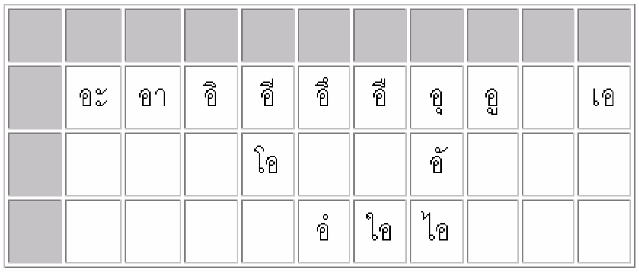
Figure 3.4 Thai consonants.

Figure 3.5 shows Thai vowels. The vowels can be divided into simple and compound vowels. The compound vowels are made up of a combination of a simple vowel with certain consonants or a simple vowel used in conjunction with another simple vowel. The vowels are composed from fourteen unique marks as shown in Figure 3.6 and three consonants: [อ], [ว] and [ย]. There are some rare vowels that are not shown in the table, such as [ฤ] and [ฦ], which are used in old documents and are practically obsolete. The Thai language has five tones. Four tone marks are used to represent the second to the fifth tones, as shown in Figure 3.7. The absence of a tone mark simply means the first or fundamental tone. Thai numerals from 0 to 9 are shown in Figure 3.8.



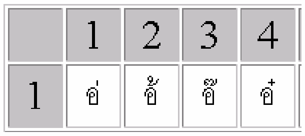
	1	2	3	4	5	6	7	8	9	10
1	อะ	อา	อิ	อี	อึ	อื	อุ	อู	เอะ	เอ
2	แอะ	แอ	โอะ	โอ	เอาะ	ออ	อัวะ	อัว	เอียะ	เอีย
3	เอือะ	เอือ	เออะ	เออ	อำ	ใอ	ไอ	เอา		

Figure 3.5 Thai vowel forms ([อ] is used as template holder).



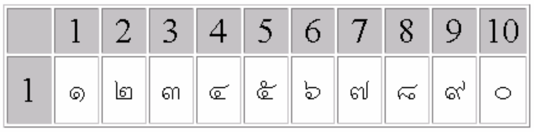
	อะ	อา	อิ	อี	อึ	อื	อุ	อู		เอ
				โอ			อั			
					อํ	ใอ	ไอ			

Figure 3.6 Unique symbols of vowel marks.



	1	2	3	4
1	อ่	อ้	อ๊	อ๋

Figure 3.7 Thai tone marks.



	1	2	3	4	5	6	7	8	9	10
1	๑	๒	๓	๔	๕	๖	๗	๘	๙	๐

Figure 3.8 Thai Numerals.

3.3 Thai writing system

Thai writing begins from left to right and from top to bottom. It is a character-based language in which one word consists of characters taken from a relatively limited pool of characters. Distinct differences between the English and Thai writing systems are as follows: English utilizes single level writing, while Thai uses multiple level writing; English uses spaces between words and sentences, while Thai writing does not require spaces between words and has no hard and fast rules for separating sentences with spaces; English has multiple forms of acceptable writing styles, e.g. cursive and block writing, while Thai is written in only one style though with more relaxed rules than in English writing. The Thai writing system has four writing levels. These levels are the tone, upper vowel, body, and lower vowel levels, as shown in Figure 3.9.

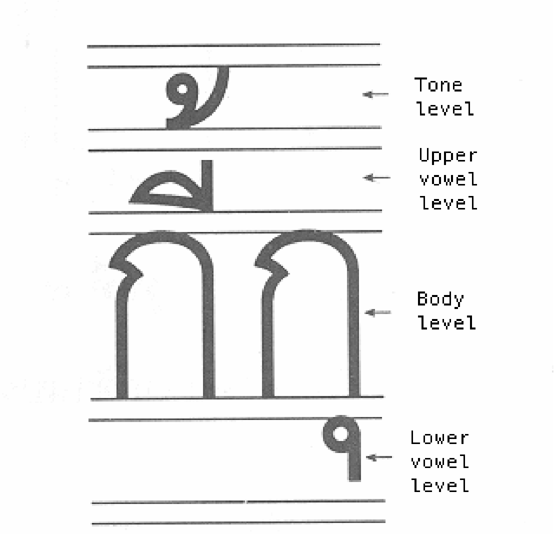


Figure 3.9 Thai language uses four writing levels.

The tone level is the highest position in a line followed by the upper vowel, the body, and the lower vowel level. The simple rules for placing characters at specific levels are: consonants are always located at the body level and the tone marks are always located at the tone level. Thai numerals are always written at the body level. The other characters have their specific positions.

The body level has the most height among other levels. Most consonants are written only in the body level. Some consonants occupy more than one level such as [ฏ] which occupies both the body and lower vowel levels, or [ป] which occupies both the body and upper vowel levels as shown in Figure 3.10.

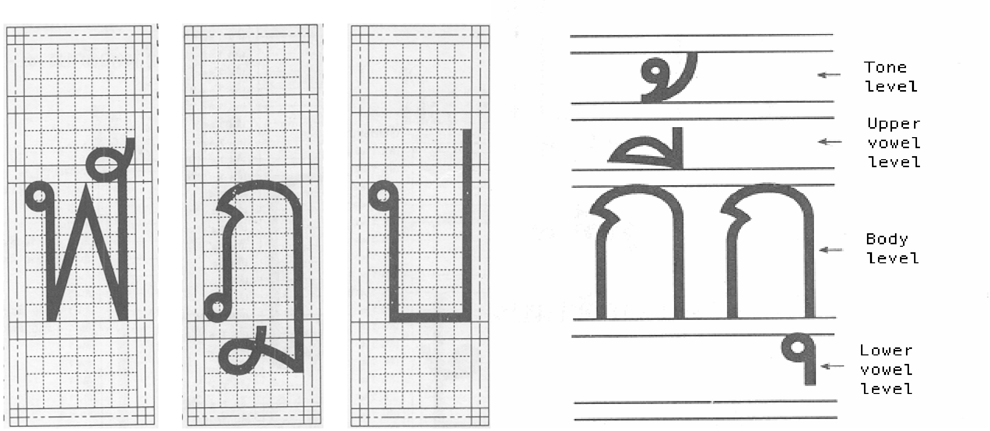


Figure 3.10 Some consonants have overlapping parts with the upper or lower levels.

Tone marks are always placed at tone level. However, it should be noted that in many printed documents and practical writings, tone marks are often found at upper vowel levels as well. Three tall vowels, i.e., [โ], [ใ] and [ไ], occupy both the body and upper vowel levels.

In the Thai writing system, when combining consonants and vowels together, they may overlap or touch each other. In most present printed documents, the positions of these combined components are adjusted to give the word a pleasing appearance. It should be noted here that there are no implementations of standards or guidelines for position adjustment in Thai character printing. The adjustments can be grouped into translation, replacement, and overlapping as shown in Figure 3.11.

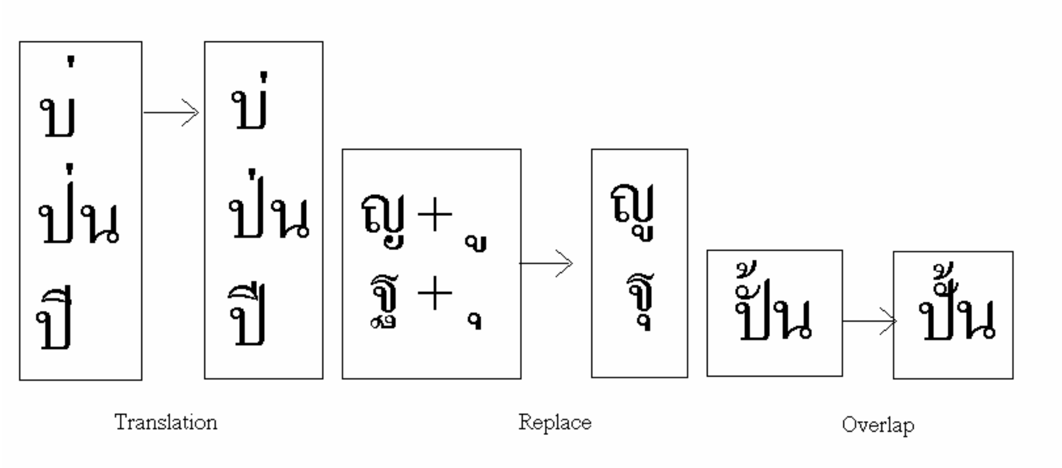
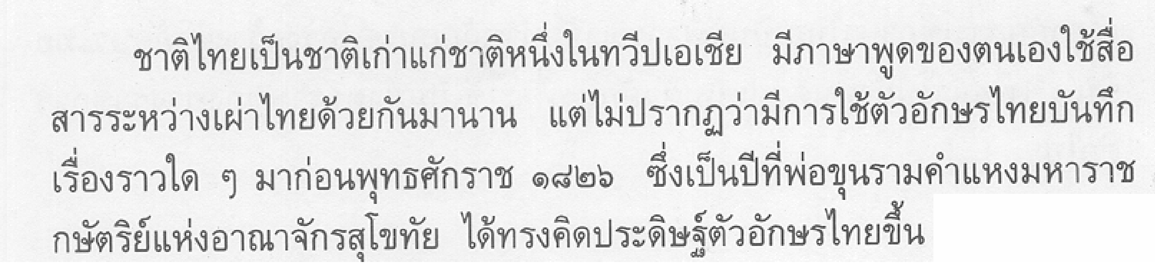


Figure 3.11 Position adjustments when letters are combined together.

In translation adjustment, vowel and tone mark positions may be adjusted to the left or the right. In replacement adjustment, based components of consonants that occupy the body and lower vowel level are replaced with lower vowels when combined together. It should be noted that replacement adjustment is required in printing but not in handwriting. In overlapping adjustment, vowels may overlap with part of a consonant that occupies the upper vowel. These modifications can be mixed together.

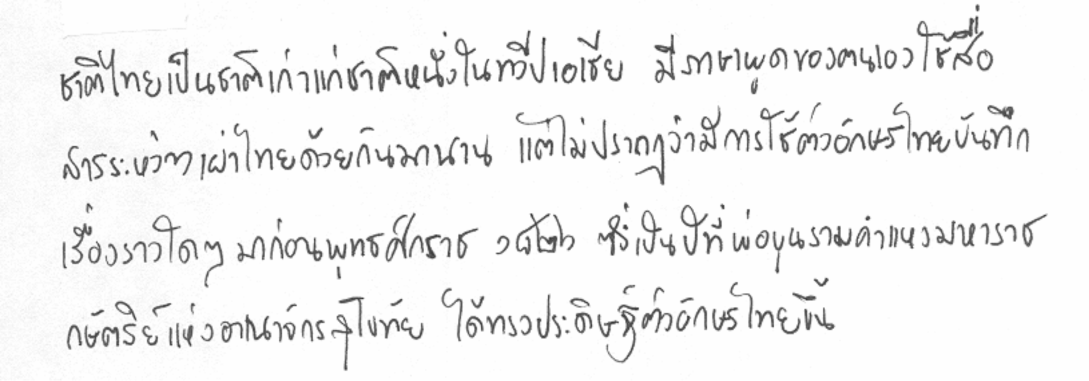
Thai words in a sentence are written together without spacing. This is one of the unique characteristics of the Thai writing system. Generally, the words will be put adjacent to previous words to make a full sentence until the entire idea of a phrase is completed. A space is mostly used for indicating the ending of an idea or expression, or the separation of a character and a numeral, e.g., "๑๘๒๖" or special characters such as character [ฯ] in the third line of Figure 3.12.

Without spaces between words, the only way to extract Thai words and sentences from a line or paragraph is to use the contextual knowledge. The first sentence in Figure 3.12 has eleven words and the second sentence has thirteen words. The translation for the first sentence is "Thai is one of the ancient nations in Asia." The translation in the second sentence is “having their own spoken language to communicate between Thai tribes for a long time.” Figure 3.13 shows an example of handwriting in running style of the paragraph in Figure 3.12.



ชาติไทยเป็นชาติเก่าแก่ชาติหนึ่งในทวีปเอเชีย มีภาษาพูดของตนเองใช้สื่อ
สารระหว่างเผ่าไทยด้วยกันมานาน แต่ไม่ปรากฏว่ามีการใช้ตัวอักษรไทยบันทึก
เรื่องราวใด ๆ มาก่อนพุทธศักราช ๑๘๒๖ ซึ่งเป็นปีที่พ่อขุนรามคำแหงมหาราช
กษัตริย์แห่งอาณาจักรสุโขทัย ได้ทรงคิดประดิษฐ์ตัวอักษรไทยขึ้น

Figure 3.12 Example of a printed Thai paragraph.

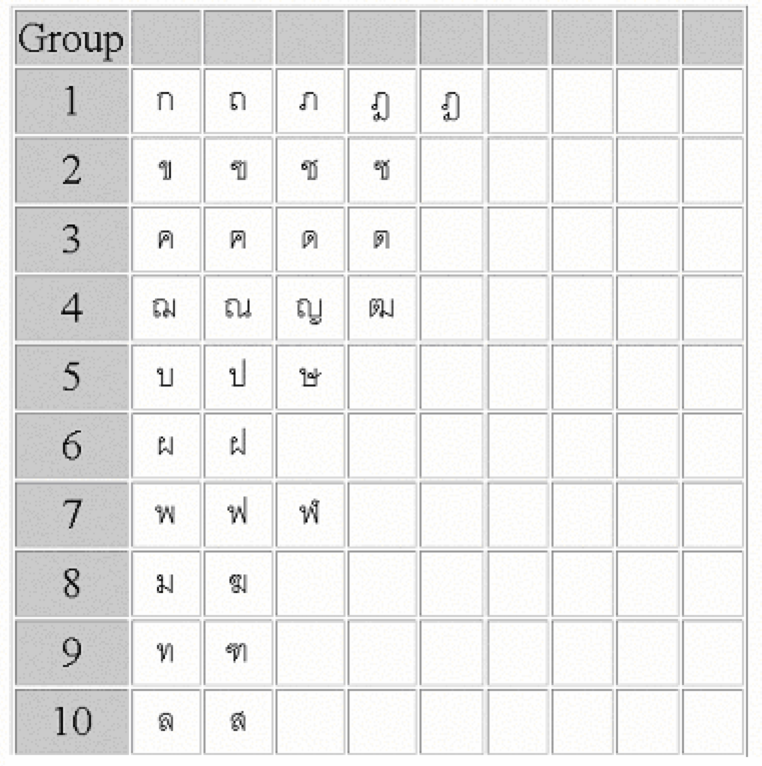


ชาติไทยเป็นชาติเก่าแก่ชาติหนึ่งในทวีปเอเชีย มีภาษาพูดของตนเองใช้สื่อ
สารระหว่างเผ่าไทยด้วยกันมานาน แต่ไม่ปรากฏว่ามีการใช้ตัวอักษรไทยบันทึก
เรื่องราวใดๆ มาก่อนพุทธศักราช ๑๘๒๖ ซึ่งเป็นปีที่พ่อขุนรามคำแหงมหาราช
กษัตริย์แห่งอาณาจักรสุโขทัย ได้ทรงคิดประดิษฐ์ตัวอักษรไทยขึ้น

Figure 3.13 Example of Thai handwriting in running style.

3.4 Thai characters’ shape

Some Thai characters are very similar. Their features are different only in some small areas. Figure 3.14 shows groups of similar characters. In each group, the similar characters are mostly separated by small marks or strokes in some position. For example, the pair of characters [ค] and [ด]. They are very similar but the shape at the beginning of character drawing or head is different.



Group									
1	ก	ถ	ภ	ฎ	ฏ				
2	ข	ฃ	ช	ซ					
3	ค	ฅ	ด	ต					
4	ฌ	ณ	ญ	ฒ					
5	บ	ป	ษ						
6	ผ	ฝ							
7	พ	ฟ	ฬ						
8	ม	ฆ							
9	ท	ฑ							
10	ล	ส							

Figure 3.14 Group of similar Thai characters based on by shapes and width.

3.5 Thai characters’ coding

The result of Thai character recognition is the codes that represent Thai characters. Currently, Thailand’s standard organization accepts the 8- bit TIS620 by Thai Industrial Standard Institute (TISI) of Ministry of Industry, Thailand. However, in decoding Thai there is a variety of coding schemes in different environments. In the Microsoft Windows operating system, for example, the code is window-640 in windows mode or code-874 in DOS mode. See Microsoft (1996) for details. For international 16-bit Unicode, Thai uses Unicode 3.1 (U+0E00 – U+0E5F) for the code-by-code mapping with the characters A0-FF of TIS 620.

3.6 Thai character recognition research

In order to understand the scopes and difficulties of the problem in Thai handwriting recognition in legal amount, previous studies related to this topic should be discussed. This will give the general idea of how the recognition systems for machine-generated and human-generated documents are developed and the current state of the research. Then the more specific details in the sub-process concerned with the unique characteristics of the Thai language are discussed in the following topics. The background information discussed is the history of developing both Thai character printed and handwritten recognition.

Kimpan (1983) is among the first pioneers in printed Thai character recognition research. His work began in 1983, with Thai characters recognition using samples from typewriters. In the earliest study of printed Thai character recognition, the two-stage recognition processes were used to reduce the size of possible candidates. Kimpan uses Karhunen-Loeve expansion to handle the features analysis of printed Thai characters. The features used in his experiment are character width and height. Most of the samples in the earliest studies were mechanic-typewriter produced. Obviously, the quality of typewritten documents are very poor compared to those of laser printer or typographic. Broken strokes, inconsistent ink, dirty hammer and inconsistent pressing weight are common for typewriter documents. These artifacts create even serious problems. The recognition of mechanic-typewriter document in Thai is very difficult compared to that of the printed handwriting.

Other research attempt to find more reliable features to identify the character classes, for example, center of gravity and histogram profiles. Most of the features are analyzed using the zoning techniques (Ratee1985; Airphaiboon et al 1994; Chatwiriya 1996). Later on, researchers of Thai character recognition change their interest to the laser printer-generated documents, instead of the typewriter-generated document because the using of laser printer becomes common in most businesses. The document generated from a laser printer is sharp and consistent which requires simple pre-processing algorithms before the recognition process. The recognition process initially works based on IF-THEN-ELSE rules to classify the character classes. The neural network, which is initially used in the Latin alphabet recognition systems, comes to be used with Thai character recognition systems (Khunasarphan & Lursinsap 1993).

There are two commercial products for Thai printed character recognition available right now by Atrium, NTS publishing, and NECTEC (Tanprasert et al. 2000). However, these products are only for the printed characters with specific fonts and sizes, and their performance lags behind those of the Latin recognition system.

For Thai handwriting recognition, Hiranvanichakorn (1984) first purposed Thai hand-printed character recognition in 1984. A set of topology code is extracted to find the character class. Hor (1985) purposed a Thai handwritten numeral recognition for a single character of 0 to 9. He uses the syntactic recognition approach by thinning the character and extracting a set of chain code and computes the distant of transformation of input model with model of the character classes. There have been only a few papers that attempted to work on handwriting recognition since then.

Khunasarphan & Lursinsap (1993) proposed the Simulated Light Sensitive (SLS) feature for Thai handwriting. These features are based on the biological study of visual cerebrum and visual cortex models. Their experiment involves a recognition system with a two-stage neural network; the first neural network classifies image into 8 groups, and then second neural network classifies the subgroup into 44 character classes, with recognition rate of 70 percent. Phokharatkul & Kimpan (1998) uses the cavity properties as their feature and use neural network as their recognition engine. Then in their other work (2000) they use the Fourier descriptor or the Fourier coefficient with neural network as the recognition engine but with different learning algorithm. The results of the last paper seem very promising with the recognition reported as 99 percent accurate. However, their scopes are limited with isolated characters.

3.7 Thai character segmentation

The character segmentation process is one of the important process in handwriting recognition systems because if characters cannot be segmented correctly, the character recognition will definitely wrong. However, Thai segmentation processes are rarely mentioned in the earlier literatures, and most of the studies only concentrate on the recognition process. The input characters are given either by manual segmentation or by computer generated.

In the early models of the mechanic Thai typewriter, characters were typed with equal space and fixed level. With the following assumption as characters are not broken, crossing or touched from near by characters, the segmentation algorithm is simply by horizontal and vertical scanning. The valleys in horizontal projection profile indicate the space between lines and the level in the same line. The valley in vertical profile projection should indicate the space between characters.

The idea to segment printed character is illustrated in Figure 3.15. There are three words in this Figure. On the right of the image, the horizontal scan show a few valleys at positions 3, 15, 20 and 40, which indicate different levels in a line of the text image. The vertical projection under the input image shows many valleys. These valleys indicate possible positions to segment characters. In the simple cases, there are two steps to segment printed characters. The characters in each writing level are separated by position indicated by two valleys in the horizontal scan. This is to reduce the interfering affect to another level. Then, the valley positions in the vertical projections can be used to separate characters from each other in each level.

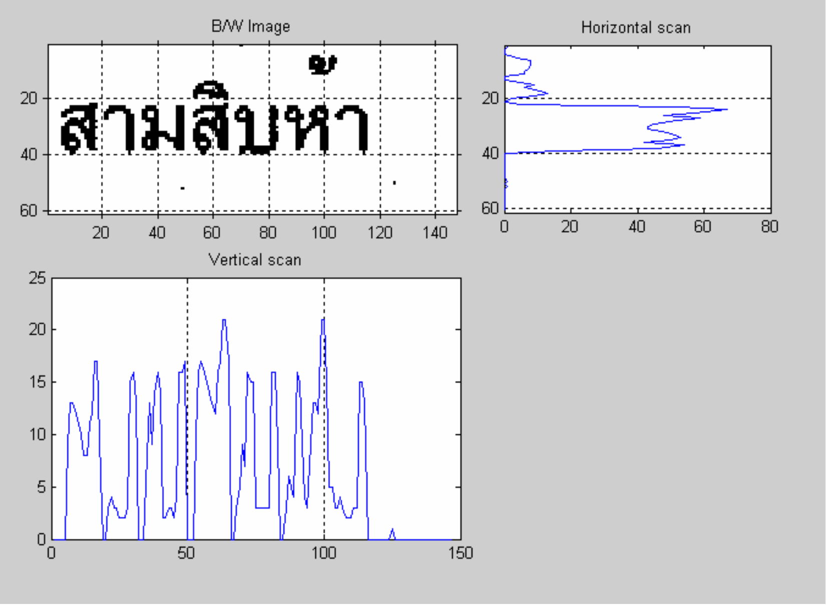


Figure 3.15 Character segmentation by vertical and horizontal projection profiles.

However, the printed character that are produced by the newer generation typewriters might have flexible positions and even overlapping or touched with each other as shown in Figure 3.16.

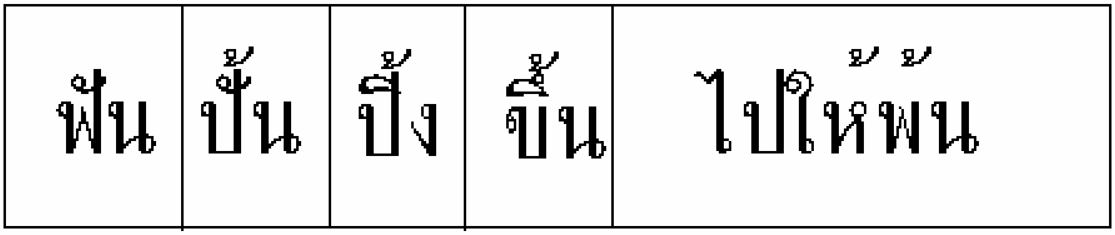


Figure 3.16 Example of printed words having connecting or overlapping characters.

For overlapped characters, contour or region separation concepts should be used. Airphaiboon (1989) enhanced the character segmentation in the study of handwritten recognition using histogram analysis and contour following. In his technique, horizontal histogram analysis will be used to indicate the consonant level. Then the vertical project will be use to locate each character from this level. However, some characters have tall structure, such as [โ], which occupies both consonant and upper vowel level. Up to this step, the extracted part is the only part in the consonant level. Overlap parts in higher or lower than consonant level can be detected by boundary detection, then contour tracing is used to retrieve those parts.

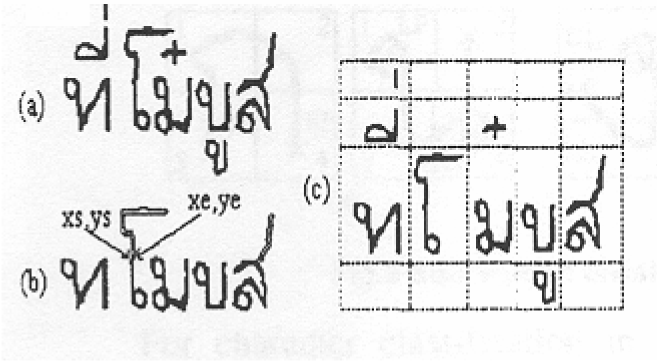


Figure 3.17 Character segmentation technique by Airphaiboon (1998).

Figure 3.17 shows details of character segmentation suggested by Airphaiboon. In Figure 3.17 (a) the image of a non-existent Thai word in which each character is isolated. By using horizontal histogram analysis, the body level of the word is detected. In Figure 3.17 (b), the body level and the detection and retrieval of the upper part of character [โ]; xs, ys and xe, ye are the positions that the upper part of character [โ] are crossed the top of body level. This top part of [โ] is retrieved by contour following algorithm. The segmentation results are shown in Figure 3.17 (c). This technique can segment character from the overlapped handwriting but not the touching and crossing handwriting. Chatwiriya (1996) suggested a solution works for the connected or crossing of nearby characters in printed Thai document by defining addition classes for the possible combinations of the connecting characters and use them as one class of character. This idea is feasible for predefined fonts system.

3.8 Features in Thai character recognition

As character recognition has been studied for a long time in English, numbers of feature both character and word level are extensively studied. These studies can be applied directly with languages that use the same alphabet. However, since the Thai alphabet differs from the Latin alphabet in many ways, these studies cannot be directly applied to the Thai language. Features used in previous studies of Thai character recognition, both printed and handwriting, are grouped into 1) structural properties 2) projection profiles 3) topology description 4) simulated light sensitivity features 5) Fourier descriptor features.

3.8.1 Structural properties

The structural properties of characters such as the width and height ratio, and center of gravity can be used to classify few groups of printed Thai character. In case of [ฒ] and [เ], these two characters are different in shape or topology too. However, there are some characters which have the same topology, such as consonant [ข], [บ] and [บ], but different in width and height ratio. This feature is very efficient in printed Thai character recognition (Kimpan 1984; Kimpan & Walairacht 1993; Thumwarin & Chittayahothorn 1998). For example, vowel [เ] has the lowest width and height ratio while consonants like [ฒ] have the highest. Figure 3.18 shows characters which can be classified by height and width ratio.

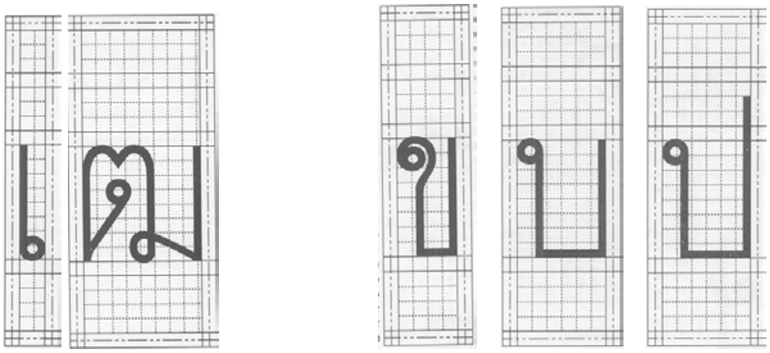
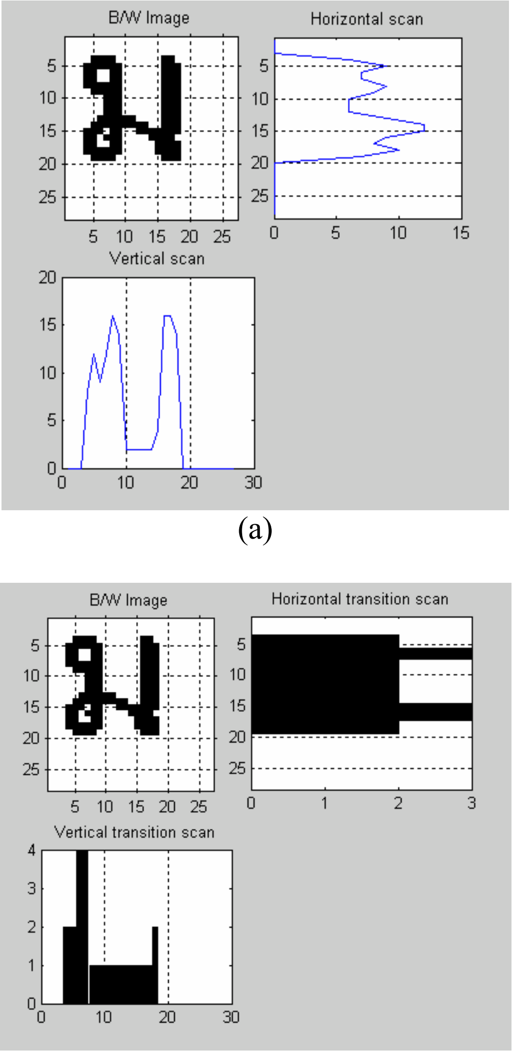


Figure 3.18 Different in height and width ratio can classify subgroup of character.

3.8.2 Projection profiles

Projection profiles are among the basic features used in pattern recognition, especially for printed character recognition. The common types of projection are horizontal and vertical. Horizontal and vertical projection profiles of character images transform two-dimensional topologies into two vectors of scan code. This vector feature can be used to classify a normalized printed character by a pattern matching process (Kimpan & Walairacht 1993; Chatwiriya 1994). The projection or scan profiles are initially referred to histogram profiles which count the number of pixels corresponding to interested values in a reference axis. The other form of projection profiles is transition profile. This projection counts the background and foreground transition in a reference axis. Figure 3.19 shows both histogram and transition profiles of the character [ม].



(a)

(b)

Figure 3.19 Projection profiles of character [ม]. (a) Histogram profiles. (b) Transition profiles.

3.8.3 Topology features

In every written language, each letter can be described in terms of composing basic structures and the essential part. Ideally, this information, if extracted, is the distinguished features that can be used to identify the unknown input character. The examples of this description are the positions of the beginning and the ending of a character and various types of concave and convex of a character's topology. To extract these features, algorithms are developed to detect these descriptions, and then quantity and quality of these features are analyzed. The studies of topology features for Thai character can be categorized into basic structure description and topology description (Airphaiboon et al. 1994 & Kijsirikul et al. 1998, 2000).

Unique structures feature is the natural way to identify a specified character for humans. By defining some unique structures, the measurement of quality and quantity of those features can be used to classify a character. For example, most Thai characters begin with the writing of a small loop, called "head" and end with a drawing called "tail". By considering the locations of head and tail, many groups of character can be classified. The head is detected by edge detection and contour following process. Thinning and line tracking algorithm detect the tail.

Airphaiboon et at (1994) describes a character structural detection in four steps: 1) a character is thinned to its skeleton structure by some thinning process; 2) a line tracking algorithm reveals the position of line termination, this is the indication of the tails of the character; 3) Edge detection algorithm reveals the edged image of a character for both the outer contour and the inner contour or loop. 4) By removing the outer contour of the character, then the inner loop, if it exists, would remain. The center of the loop will be the location of the head, as illustrated in Figure 3.18.

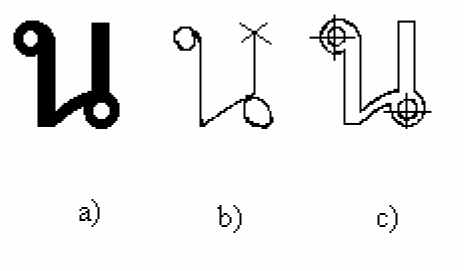


Figure 3.19 Head and end line detection. (a) Original characters. (b) Skeleton pattern. (c) Edge pattern. Reproduced from Airphaiboon (1994)

Standard topology is another form of character shape description. Cavity topology is one of the interesting structures studies in Thai handwriting recognition. Cavity defined as a region of points bounded by the stroke on at least three sides. Convex and concave cavity feature are used in recognizing Thai hand-printed by (Hiranvanichakorn et al 1984). Phokharatkul & Kimpan (1998) defined directional and topology cavity include north, south, east, west, center, and hole. These cavity features are then used in a character handwriting recognition system as shown in Figure 3.20.

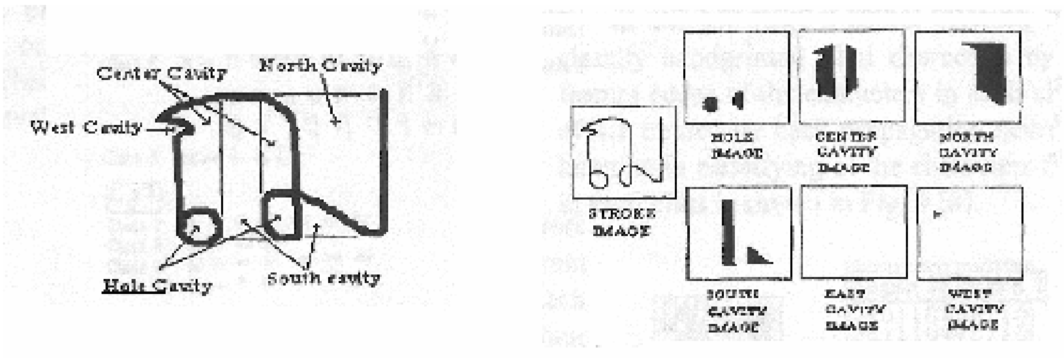


Figure 3.20 Cavity features, reproduced from Phokharakul & Kimpan (1998)

Zoning is a simple way of topology description. It divides an image into zone or sub-region of a grid system. Observing the features according to specified zone bring information that is more meaningful. Feature extraction combined with zoning technique appears in most of the papers in Thai character recognition because many Thai characters are distinguished only by some specified area or zone. For example, character [ด] and [ต] are very similar but the cavity type in the top area of both characters are different. If the cavity in this zone can be extracted, it can be used to identify these characters. Most of the feature extraction methods for Thai character image include zoning in the analysis. Ratee (1985), Chatwiriya (1994), Choruengwiwat (1998); Kimpan & Walairacht (1993) all used center of gravity as the center of the template. Tanprasert & Sae Tang (1999) used the head of the character as the reference to some pre-designed template. Figure 3.21 shows head and tail feature extraction and the zoning for Thai handwriting recognition.

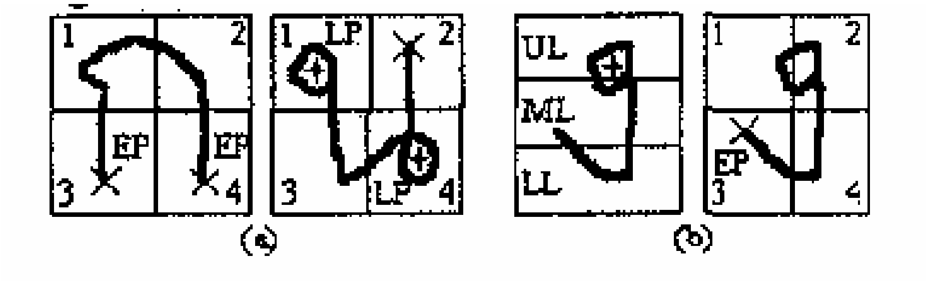
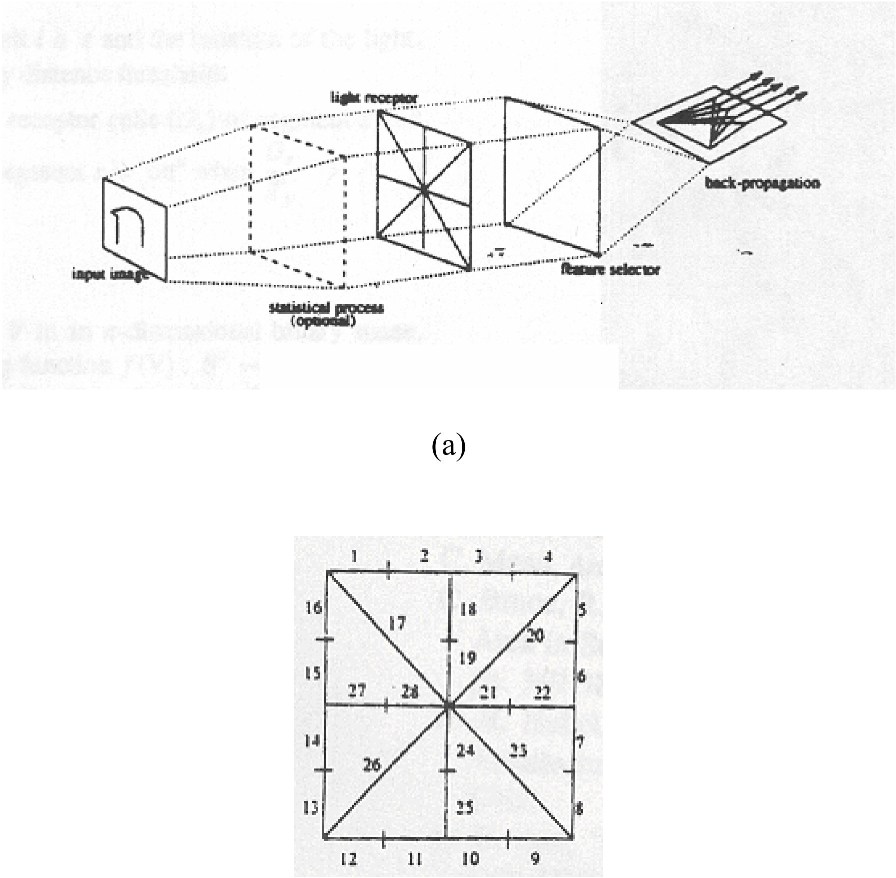


Figure 3.21 Zoning and feature of interest (head and tail), reproduced from Airphaiboon (1994)

3.8.4 Simulated Light Sensitive

SLS (Simulated Light Sensitive) feature is proposed by Khunasaraphan and Lursinsap (1993) to recognize Thai characters. In their recognition model, a receptor consists of an array of NxN cells. Each cell responds to the different amount of intensity and grating orientation of light fallen on its surface. The character image is projected onto a field of sensor segments where each segment makes up a number of sensor elements. The state of a given sensor element is either on or off, depending on the light intensity. If the neighboring region projected with intensity greater than some threshold value, the sensor state is on. The state of a sensor segment depends upon the state of a majority of its sensor elements. This feature is used with handwritten recognition system with a neural network recognizer (Khunasaraphan & Lursinsap 1993; De Vel et al 1995). Figure 3.22 shows the structure of a recognition system using the SLS features.



(a)

(b)

Figure 3.22 Simulated light sensitivity features. (a) Structure of light receptor element. (b) Receptor cell. Reproduced from (Khunasaraphan & Lursinsap 1993).

3.8.5 Fourier descriptors

Fourier descriptors are used extensively in studies of the Latin alphabet but it has only recently received attention from Thai researchers. Phokharatkul and Kimpan (2000) used the Fourier descriptors with Thai handwritten recognition with a successful result. First, contour following or edge detection algorithm finds the contour of the binary image. Then, the Fourier descriptor coefficients are computed from these contour points. The original image can be constructed from all of these coefficients. However, only first few order of these coefficients can be used to represent the essential or main structure of the character. Figure 3.23 shows the original binary image and its contour or edged image. Figure 3.24 show the contour image and its contour points of 78 points. Figure 3.25 shows images constructed with a different number of coefficients from 6 to 8. Phokharatkul and Kimpan (2000) report the effective number of the coefficient is approximately 10-12 to describe a Thai character.

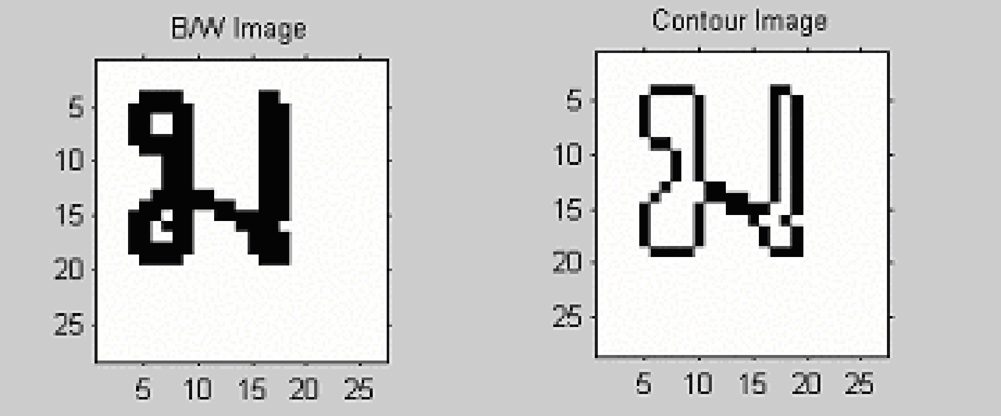


Figure 3.23 Binary image and its contour image.

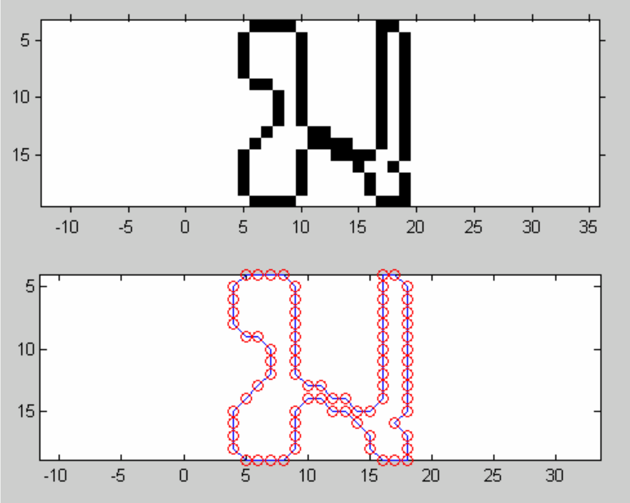


Figure 3.24 Contour edge and edge points.

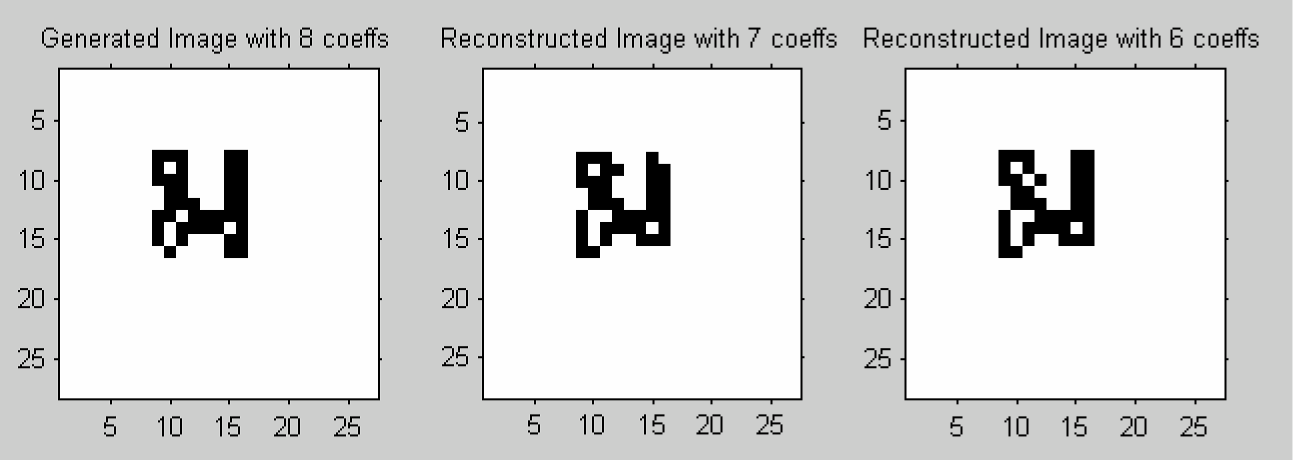


Figure 3.25 Reconstructed images from 8 to 6 coefficients.

3.9 Recognition approaches

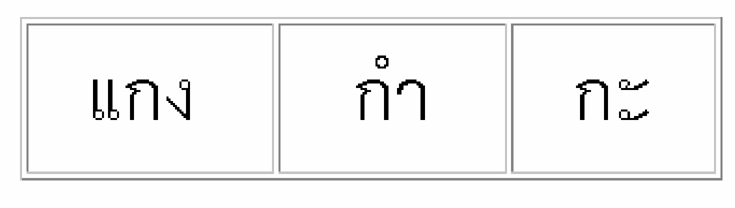
Two approaches of recognition for Thai character are multi-stages and single-stage recognition. The first approach will do rough classification into a group of similar characters then refine in the next stage to identify the character class. It works very well with quite a large set of symbols and having similar character. The ID3 recognition is an example of this approach (Thumwarin & Chittayasothorn 1998). The second approach, input will be classified to a specific class of character in one step. The multiple-stages neural network classifier of Tanprasert (1996) is an example of the second approach.

Currently, Thai documents, such as newspapers or textbooks, are using a significant number of English alphabets as the references to borrowing-words or words translated from English. Arabic numerals are commonly found instead of the original Thai numerals. Therefore, some studies suggested that the practical Thai recognition system should include the Latin alphabet as well (Tanprasert et al 1993; Tanprasert et al 1996).

3.10 Postprocessing

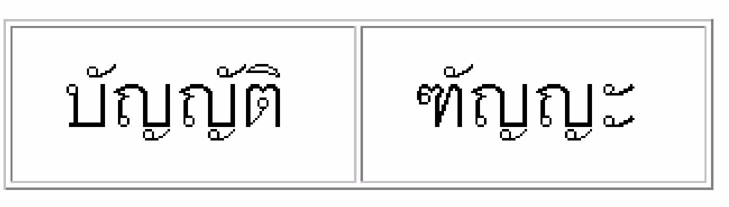
For character recognition system, the inputs to the recognition module are features vectors of an unknown character. The recognition result is interpreted from the information obtained from one character or one sub-character towards existing class models. If the information provided from a segmented unit is enough, then the correct class can be identified. However, sometime only information extracted from one segmented unit is not adequate.

Some Thai characters consist of multiple isolated components. For example, vowel [แ] has two components. Each component looks like vowel [เ]. When the word image is analyzed to segment the single characters, these components are extracted instead. Assuming the recognition module can correctly identify these components. They still need to be process further as one valid class. The recognition of two consecutive [เ] should result into one [แ]. These exceptional cases would require a post-processing after recognition. However, very few of studies mention this necessary process.



แกง	กำ	กะ

Figure 3.26 Example of Thai words contain vowels which have multiple separated components.



บัญญัติ	ฑัญญะ

Figure 3.27 The mark [ั] at different levels represents different vowel and part of consonant.

3.11 Current Thai handwriting recognition systems

According to the literature, Thai writing recognition research is at the beginning state and very few. The research experiments are conducted with different set of data and collecting environments. The recognition systems concentrate at character level, and require isolated character writing, i.e., no touching or crossing characters. The character classes in the experiment are different. Most experiments include forty-four consonants, (some use forty-two classes because two classes are obsolete), and selection of ten Thai numerals, seventeen vowel, four tone marks and special marks, ten Arabic numerals. Nevertheless, direct comparison between these studies is not sensible because their experiments were conducted with different test sets and conditions. Up to the present, there is no common of test databases either for both printed or handwriting in Thai. Table 3.1 provides summary of techniques and their performance.

Author	Descriptions	Year	Char. Class	Perfor-mance %(c/m/r)	no. of writers	Test set
Phokharatkul & Kimpan	Recognize rotated and scaling characters using Fourier descriptor of character boundary. Contour of specific zone are used to separate similar character. Use genetic NN.	2000	63	99.12 0.23 0.65	60	13,500 characters of 1,200 unspecified words
Veerathannabutr & Homma	Normalization process using evaluation of Variation Entropy. Use modified Fuzzy Direction Code Histogram (FDCH) & Zoning (39 features), collect data from tablet.	2000	81	>90	1	Training: 3,240 Validation : 1,620 Test set: 3,240
De vel et al.	Proposed a two-stage classification procedure using backpropagation multilayer classifier and discriminant analysis technique, simulated light sensitive. Feature vector are reduced into 32 features using Daubechies wavelet transformation matrix.	1995	41	85.0 to 98.7	1	N/A
.Methasate et al	Purpose set of fuzzy features descriptor with zoning and use fuzzy-syntactic decision. Use two examples in two pairs of similar characters.	2000	N/A	N/A	N/A	N/A

Table 3.1 Summary of Thai handwriting character recognition research.

(c/m/r = percent of correct rate/mis-classify rate/reject rate).

Author	Descriptions	Year	Char. Class	Perfor-mance %(c/m/r)	no. of writers	Test set
Airphaiboon et al	Character recognition from word They proposed a line (main body level) approximation, character segmentation based on histogram and contour following. Decision tree Feature: Topology (End point, Cross point, Loop) combined with 3 zoning types; feature code, loop type, sub regions, and width / height ratio	1994	64	99.0	10	100 isolated word (character freq omitted)
Khunasarphan & Lursinsap	Proposed Simulated Light Sensitive Model as a feature based on the biological study of models of visual cerebrum and visual cortex. Applied to 44 consonants. Use two-stage NN. First NN classify image into 8 groups, second NN classify the subgroup to 44 characters.	1993	44	70	N/A	No details, Set: 77 samples (no details on char freq)
Choruengwiwat et al.	Propose recognition of Thai handwritten using modify stroke changing sequence (SCS).Modified SCS features detected at head of Thai characters can be used to classify character classes by if-then decision based.	1998	44	92	28	5 copies of 44 consonants

Table 3.1 Summary of Thai handwriting character recognition research (continue).

The experiment by Phokharatkul & Kimpan and Airphaiboon et al uses hand-printed sample words from number of writers (ten to sixteen) and shows very impressive results, higher than 99 percent correct recognition rate. However, some experiments allow unspecified words, therefore the distribution of character classes of test samples is not equal. Most of the experiments, this information is often omitted. Doubt remains concerning how the accuracy of the recognition rate is measured.

Most of the research concern with Thai printed characters recognition. Only few research interested in handwriting characters recognition problem. For handwriting recognition studies, the character separation processes are performed based on the contour following algorithm. There is no study for word separation in Thai recognition system. The features for recognition are also studies only at level of character. There is no report dealing with the recognition at the word level for Thai handwriting. Since the research is conducted with different sets of data and assumptions, the results are not comparable. The best result of Thai character recognition system comes from Phokharatkul, which use Fourier descriptor as their features and uses the neural network as the recognition engine.

3.12 Summary

In this Chapter Thai characters handwriting and recognition systems are described. A brief history of Thai script is given. The basic knowledge about the Thai alphabet and writing system are illustrated. Thai alphabets are similar in shape and require some practice to distinguish each character, even in printed form. The writing system which has four writing levels, the different position of each alphabet and no boundaries between words make the Thai writing system different and therefore even harder to implement a recognition system compared to that of the Latin alphabet. The more difficult problem comes from the relaxed writing style of Thai language which results in touched and overlapped strokes.

In conclusion, of the review of Thai handwriting system, the researches of Thai handwriting system are at the beginning state. The previous studies in Thai handwriting recognition system are limited at the character level and with constraints. The results of these studies are not reliable because of the small number of writers. Many experiments did not provide mandatory information in the experiment such as distribution of characters in the test. Moreover, comparing these results numerically are not sensible, because their different data set and assumptions. Phokharatkul & Kimpan (2000) seem to be the most successful group, with the result of 99 percent from 60 writers. Their study is limited to unconnected character writing or recognition at the character level. In this study, the problems which occur with the connected writing of characters are the main concerned. This is considered as recognition at word or clause level. However, in the model of the proposal, the recognition at character level is one of the key-systems, which can benefit from Phokharatkul & Kimpan (2000) studies. However, the other parts of the proposed system require new explorations or novel concepts.

Chapter 4

Thai Handwriting Data Collecting and Analysis

In this Chapter, the words and characters used in Thai legal amounts are introduced. Then the data collecting of the Thai handwriting legal amounts is described. The data set includes handwriting characters, words and amounts. The characters and words are used mainly to implement the character and word recognition systems. The legal amounts are used to test overall system performance. Afterward, the basic information of Thai characters drawing and problematic characteristics of handwriting, which are the essential parts in designing robust recognition systems, are illustrated. However, the explanations in this study are only the surface of the real handwriting recognition complexity. Finally, some preliminary experiments are set up to evaluate how the features used previously for Thai printed characters perform with the handwriting data.

4.1 Introduction

The data collecting is the first step in the pattern recognition analysis. This process requires a set of samples representing the general population of data. In this study, Thai handwriting in legal amounts is the target data set. There are three types of handwriting which volunteers were asked to write: legal amount phrases, characters and words used in legal amount phrases. Note that the characters and words were asked to be written separately from legal amount phrases. The character and word data sets are used to implement the character and word recognition systems and the legal amount data set is used to test the system performance.

To implement the recognition system, the samples are often separated into two groups: the training and the testing group. The training data set is used to create or train classifier systems to adapt or learn to generate output with respect to pre-defined values. The systems are expected to have the capability to identify data that did not exist in the training set. The testing data set are used to test if the system can recognize the data outside their training set.

4.2 Legal amount lexicon

The legal amounts are phrases used to describe numeric values. They are often found in financial documents where the value verification is necessary. Some of the Thai words in legal amounts are the one-to-one mapping of numeric to word. The full amount is the result of joining these words. The syntax rules are used to control how these words are joined. There are seventeen words in the Thai legal amount lexicon. The words can be divided into the number word group and the tenth-power word group.

There are eleven words in the number words group for numeric 1 to 9. The mapping of the words to the numeric are the following: หนึ่ง for 1, สอง for 2, สาม for 3, สี่ for 4, ห้า for 5, หก for 6, เจ็ด for 7, แปด for 8 and เก้า for 9. There are two additional words, เอ็ด and ยี่ used for numeric 1 and 2 on the specific positions. The word เอ็ด is used when the numeric 1 is the last digit of the amount, if that amount has at least two digits. The word ยี่ is used when the numeric 2 is the second digit of the amount. There are six words in the second group for the *n*-th power of ten values; สิบ for 10, ร้อย for 100, พัน for 1,000, หมื่น for 10,000, แสน for 100,000 and ล้าน for 1,000,000. These words are grouped as tenth-power word group. These words are used after the number word group, for example, หนึ่งร้อย, สองพัน and แปดล้าน. Table 4.1 shows the amount words and their numeric value. The pronunciation is given for easy reference.

Number words group			*Tenth-power* words group		
Thai script	Pronounce	Numeric	Thai script	Pronounce	Numeric
หนึ่ง	*Neung*	1	สิบ	*Sip*	_0
สอง	*Suang*	2	ร้อย	*Roy*	_00
สาม	*Sarm*	3	พัน	*Pun*	_,000
สี่	*See*	4	หมื่น	*Murn*	_0,000
ห้า	*Hah*	5	แสน	*Saan*	_00,000
หก	*Hok*	6	ล้าน	*Larn*	_,000,000
เจ็ด	*Chjet*	7			
แปด	*Bpaat*	8			
เก้า	*Gkow*	9			
เอ็ด	*Et*	_1			
ยี่	*Yee*	_2x			

Table 4.1 The amount word, their pronunciations and value.

The amount words are made of twenty-five characters. There are fifteen consonants and nine vowels. The consonants have their base on the lower writing line, while most vowels, [ึ], [่], [ี], [ั], [็], [ั] and [้] are written above the consonants. There are two vowels, [เ] and [า], which are written at the same level as the consonants. The character set does not contain the vowel [แ] because it is exactly the composition of two vowel [เ] symbols. Table 4.2 shows all characters and the amount words of which they are composed.

Class number			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Character			ห	น	ึ	่	ง	ส	อ	า	ม	ี	้	ก	เ	จ	็	ด	ป	ิ	บ	ย	ร	พ	ั	ื	ล
Numeric	Word																										
1	หนึ่ง		x	x	x	x	x																				
2	สอง						x	x	x																		
3	สาม							x		x	x																
4	สี่					x		x				x															
5	ห้า		x							x			x														
6	หก		x											x													
7	เจ็ด														x	x	x	x									
8	แปด														x x			x	x								
9	เก้า									x			x	x	x												
_0	สิบ							x												x	x						
_1	เอ็ด								x						x		x	x									
_2x	ยี่					x						x										x					
_00	ร้อย								x				x									x	x				
_,000	พัน			x																				x	x		
_0,000	หมื่น		x	x		x					x															x	
_00,000	แสน			x				x							x x												
_,000,000	ล้าน			x						x			x														x

Table 4.2 Thai characters in amount words.

4.3 Thai legal amount handwriting data set

Collecting data is a crucial process in creating a general model for pattern recognition applications. The recognition process requires a large set of data to represent the general population of the problem. The handwriting data set in this study consists of characters, words, and amounts phrases.

The characters and words are mainly used to implement the character and word recognition systems. The legal amounts are used to test overall system performance. Characters and words in the amount lexicon are extracted from designed forms and stored in the database. The testing group is the legal amounts phrases representing the general amounts up to seven digits with varying word lengths to test the performance of the recognition system. The diagram of the whole data set is shown in Figure 4.1.

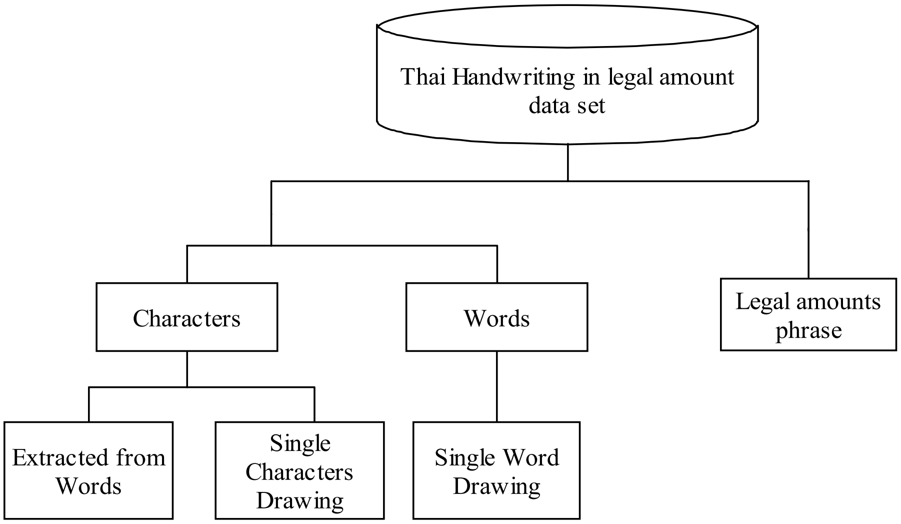


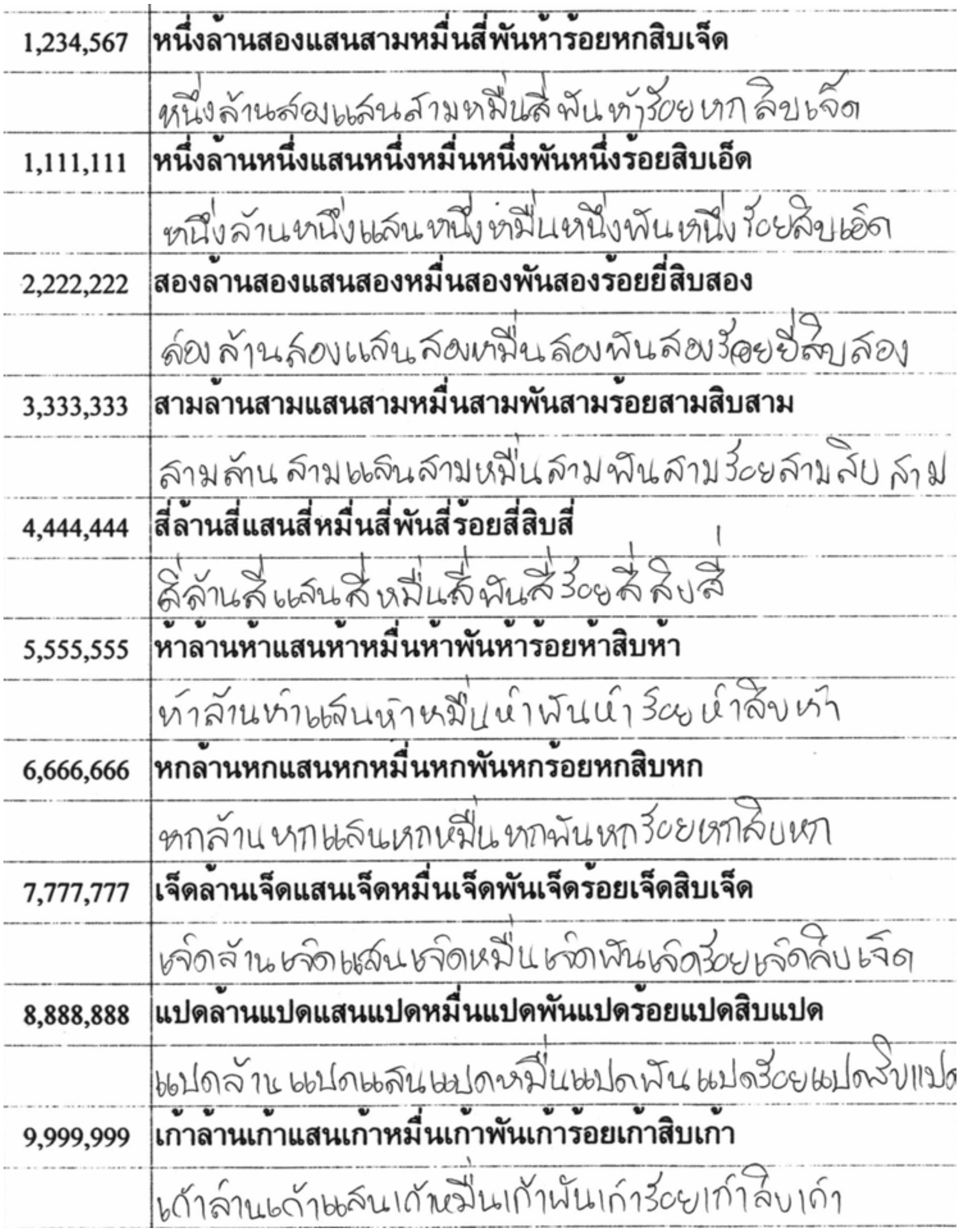
Figure 4.1Thai handwriting legal amount data set diagram.

The common reference data sets for English handwriting research are as large as a thousands writers with tens of thousand samples per class. At the time of this study, no such databases for Thai handwriting existed. Building a Thai handwriting data set with the same quality of the English data sets requires a significant amount of time and other resources which are beyond the available resources of this study. Therefore, an ad hoc database of Thai handwriting legal amount samples are designed and collected while the study is conducted in the United States. The number of writers of this data set is not large enough to claim the representation of a general population of Thai handwriting, but it could be regarded as a preliminary data set to conduct the research.

4.4 Handwriting characters data set

The handwriting characters data set is the collection of the binary images of handwriting characters used in the legal amount words. These samples are the essential parts to create the handwriting character recognition system. There are two approaches to collect the character samples: extraction from the actual amounts and from the set of single character drawing.

For the first approach, characters are extracted from the actual writing of the legal amounts. The shapes of characters are often blended to the pattern of the whole word. In other words, the ending position of the previous character affects the beginning position and direction of the next characters. Extracting characters from the actual written words requires a set of legal amount containing all words in the lexicon and having repetitive patterns. The prototype amount words are designed to be a set of legal amounts of which each amount is a seven-digit amount. There are ten amounts in the form. All amounts are seven-digit or in millions. The first amount is the legal amount of 1,234,567. The second is the legal amount of 1,111,111. The third is the legal amount of 2,222,222 and so on through 9,999,999. Figure 4.2 shows one page of Thai handwriting in legal amounts. The samples are collected from 42 people.



1,234,567	**หนึ่งล้านสองแสนสามหมื่นสี่พันห้าร้อยหกสิบเจ็ด**
	หนึ่งล้านสองแสนสามหมื่นสี่พันห้าร้อยหกสิบเจ็ด
1,111,111	**หนึ่งล้านหนึ่งแสนหนึ่งหมื่นหนึ่งพันหนึ่งร้อยสิบเอ็ด**
	หนึ่งล้านหนึ่งแสนหนึ่งหมื่นหนึ่งพันหนึ่งร้อยสิบเอ็ด
2,222,222	**สองล้านสองแสนสองหมื่นสองพันสองร้อยยี่สิบสอง**
	สองล้านสองแสนสองหมื่นสองพันสองร้อยยี่สิบสอง
3,333,333	**สามล้านสามแสนสามหมื่นสามพันสามร้อยสามสิบสาม**
	สามล้านสามแสนสามหมื่นสามพันสามร้อยสามสิบ สาม
4,444,444	**สี่ล้านสี่แสนสี่หมื่นสี่พันสี่ร้อยสี่สิบสี่**
	สี่ล้านสี่แสนสี่หมื่นสี่พันสี่ร้อยสี่สิบสี่
5,555,555	**ห้าล้านห้าแสนห้าหมื่นห้าพันห้าร้อยห้าสิบห้า**
	ห้าล้านห้าแสนห้าหมื่นห้าพันห้าร้อยห้าสิบห้า
6,666,666	**หกล้านหกแสนหกหมื่นหกพันหกร้อยหกสิบหก**
	หกล้านหกแสนหกหมื่นหกพันหกร้อยหกสิบหก
7,777,777	**เจ็ดล้านเจ็ดแสนเจ็ดหมื่นเจ็ดพันเจ็ดร้อยเจ็ดสิบเจ็ด**
	เจ็ดล้านเจ็ดแสนเจ็ดหมื่นเจ็ดพันเจ็ดร้อยเจ็ดสิบเจ็ด
8,888,888	**แปดล้านแปดแสนแปดหมื่นแปดพันแปดร้อยแปดสิบแปด**
	แปดล้านแปดแสนแปดหมื่นแปดพันแปดร้อยแปดสิบแปด
9,999,999	**เก้าล้านเก้าแสนเก้าหมื่นเก้าพันเก้าร้อยเก้าสิบเก้า**
	เก้าล้านเก้าแสนเก้าหมื่นเก้าพันเก้าร้อยเก้าสิบเก้า

Figure 4.2 Thai handwriting in legal amounts.

Since the volunteers are asked to write the amount words as usual, some characters might touch with others, or some strokes might be broken because the writing styles or use of poor quality ink. Each component might not be a single character. Figure 4.3 shows one line of handwriting in legal amount. This handwriting consists of not only the normal characters but also of the connected components, the broken parts components and the ambiguous characters.

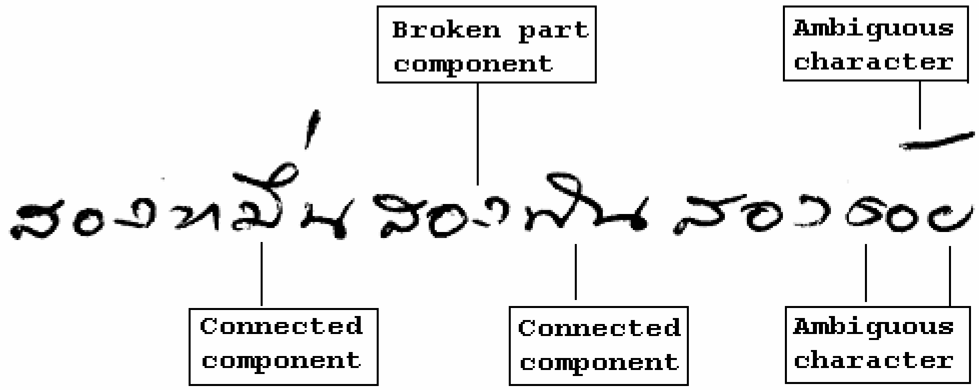


Figure 4.3 A sample of handwriting.

The automatic identification of the anomaly components is not possible. Therefore, a graphical interface is created to manually extract these components and specify their classes. An amount word image is read one line at a time, and then displayed on screen. The 256-gray level image is converted into a binary format and analyzed to find all isolated components. The system then shows the image of each component and waits for the user to identify the class. Each component can be identified by choosing image patterns that match with that component. The horizontal bar is used to select each isolated component, either character or non-character. The shape and position of each component is shown in a block with their number and marked in the line image. The interface is shown in Figure 4.4.

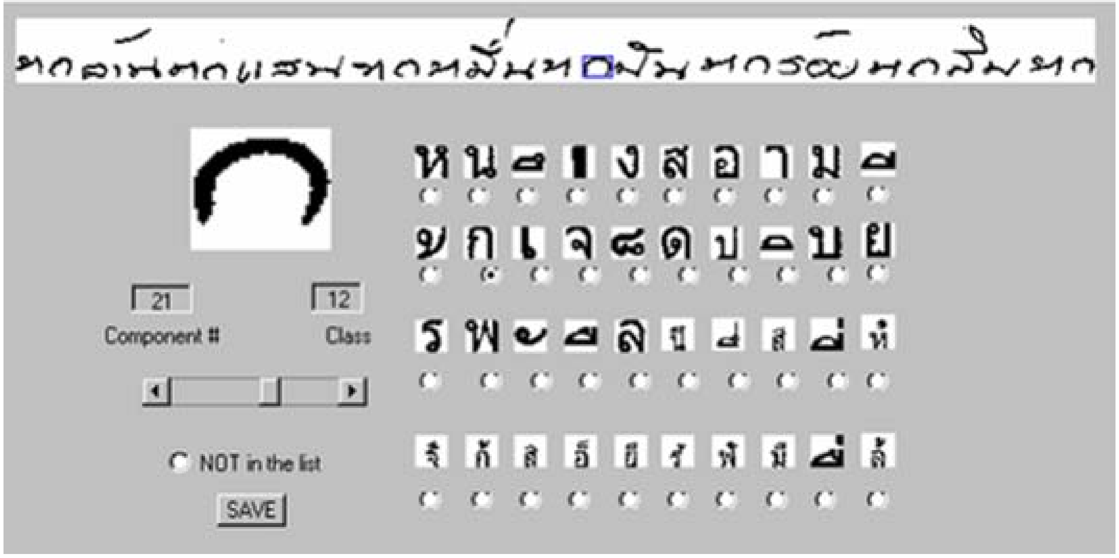
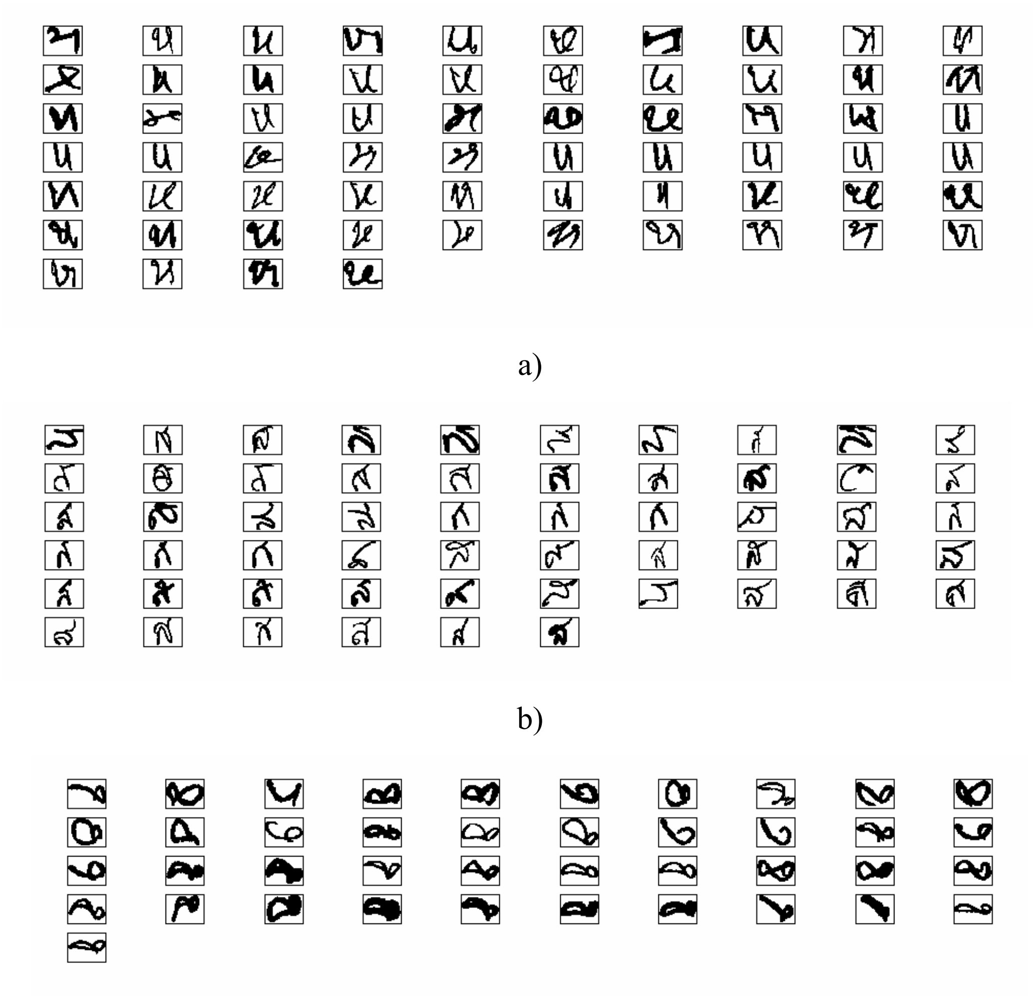


Figure 4.4 The interface of the character extraction tool.

Only the connected characters that result from the joining between the consonant and the vowel on the same vertical position are assigned to the extended class identification. The other components are assigned to the *not-assigned* class by clicking on the "*Not on the list*" button in the extraction tool interface. The connected components can be used to train the recognition system to learn the extended classes of handwriting components.

Many characters extracted from the legal amount are of ambiguous characters, i.e., characters are drawn as scribble or drawn in a unique way divergent from the standard form so they cannot be identified by other readers. These characters can be identified only when the readers consider the additional information from the whole word of that character and the information from syntax analysis. Even these characters are the real handwriting samples but they should be excluded from the training set because they can confuse the training algorithm. Figure 4.5 shows some samples of three characters extracted from the amount words.



a)

b)

Figure 4.5 Handwriting character samples. (a) Consonant [ห]. (b) Consonant [ส] and (c) Vowel [ิ].

For the second approach, the characters are extracted from the sample of separated and repeated drawing characters. The characters collected from this approach show clearer feature details than the first approach but the size and position, especially of the vowels, are found different from those drawn in actual words. To collect the handwriting characters drawn separately, a form of all characters in the legal amount lexicon are designed to be drawn with each character bounded by a significant space.

Since all characters are drawn separately, there is no problem in implementing a program to segment all isolated components from the document image at the specific positions. Figure 4.6 shows one page of handwriting characters form. This data is used as the complement set due to the reason that the first approach has the uneven class distribution and some classes are very small. The data taken from this approach is even and can be used as the compliment to the first data set. The data are taken from a smaller group of 11 Thai students at West Virginia University.

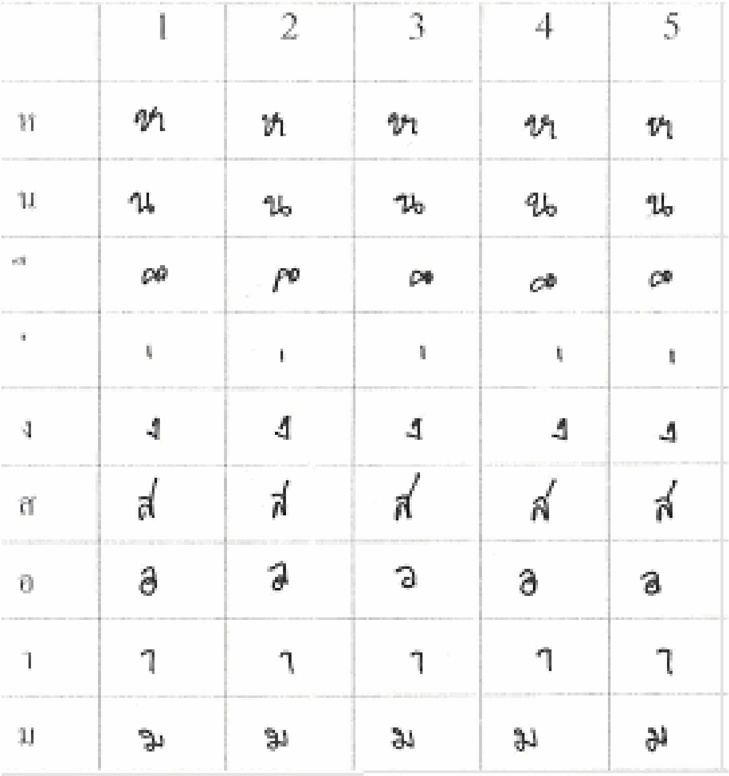


Figure 4.6 The sample of handwriting of all characters written separately.

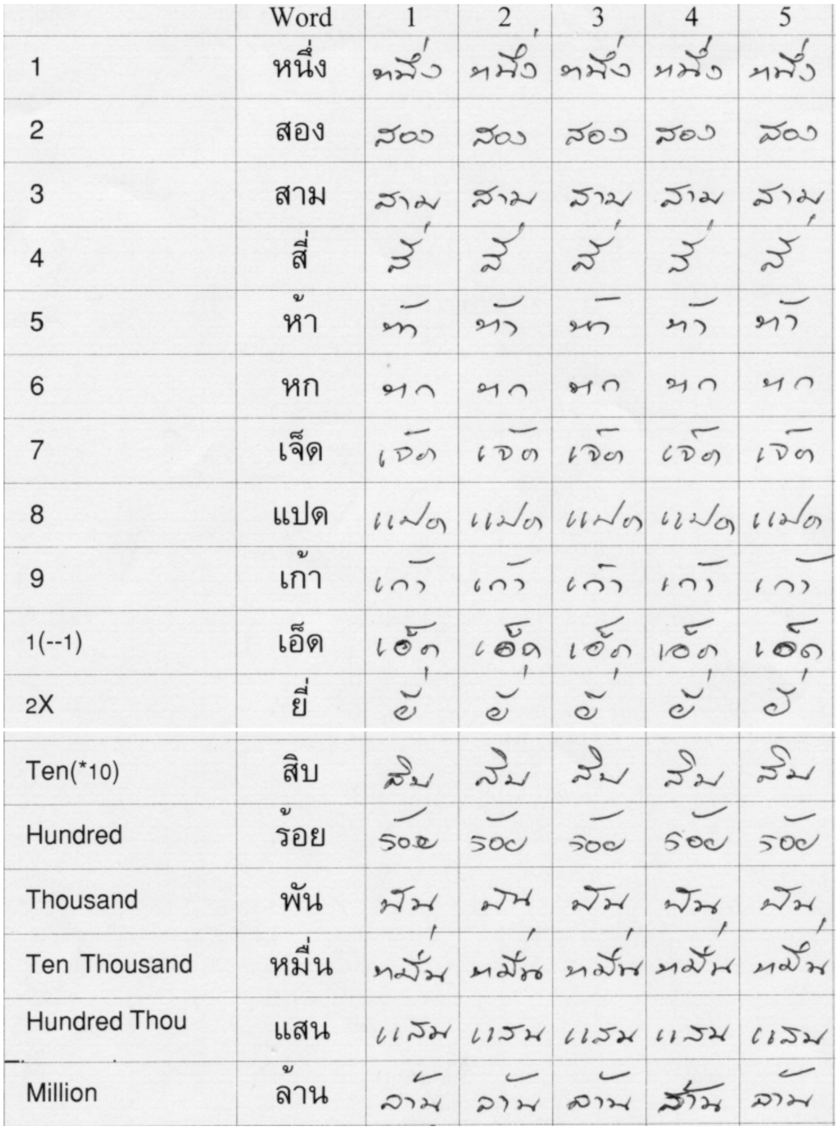
The handwriting character samples extracted from words and those which are drawn separately are accumulated, totaling 11,655 characters. The distribution of these characters is shown in Table 4.3. The characters distribution in each class is not equal because many ambiguous characters are rejected.

Character	Frequency	Character	Frequency	Character	Frequency
ห	1195	ี	235	บ	370
น	1499	้	567	ย	417
ึ	169	ก	625	ร	275
่	758	เ	2171	พ	240
ง	421	จ	297	ั	210
ส	1078	็	198	ื	109
อ	554	ด	654	ล	437
า	1228	ป	392		
ม	602	ิ	259		

Table 4.3 Character distribution in the database.

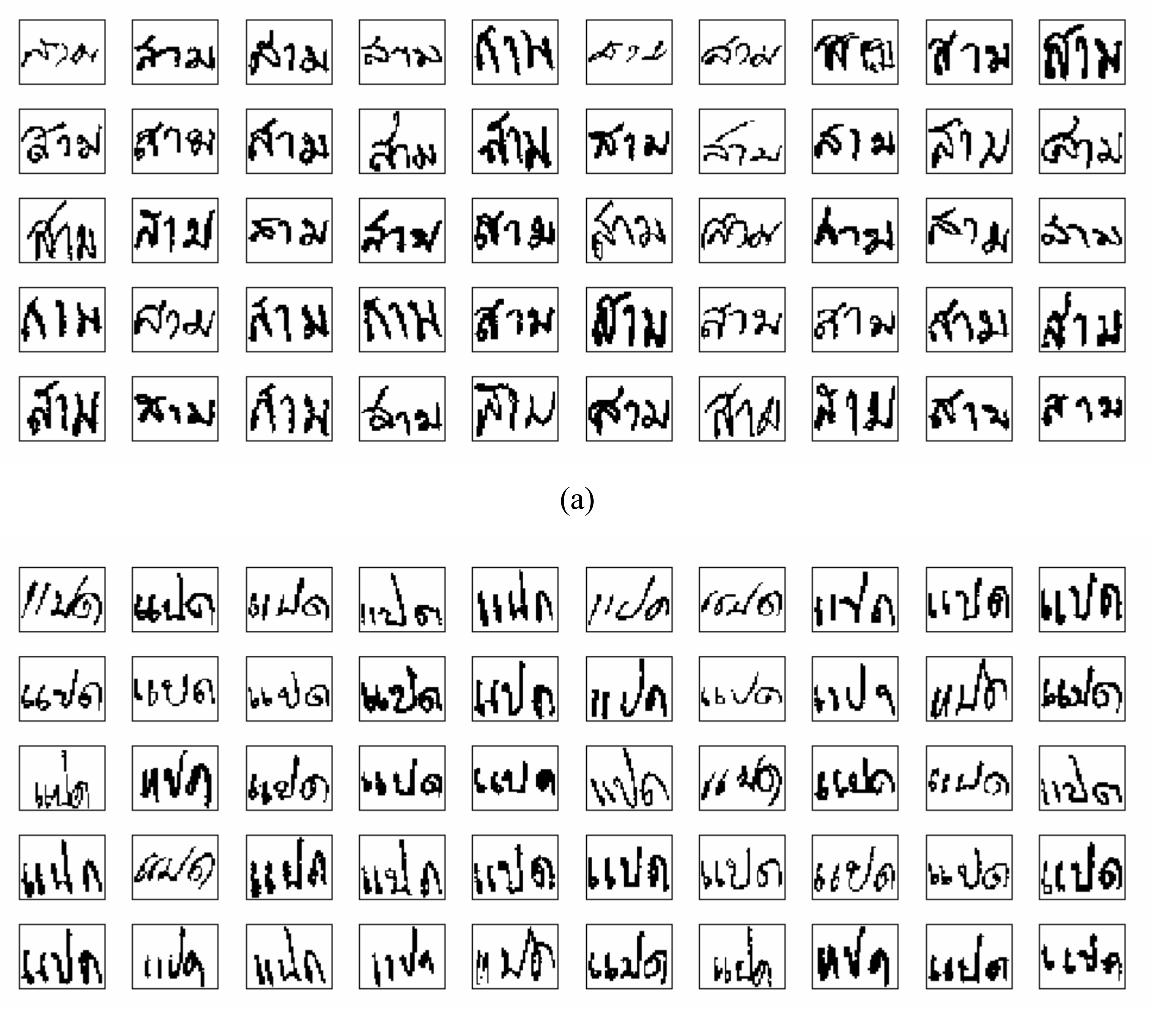
4.5 Handwriting words data set

The second set of data is the images of handwriting words. These data are required for the word recognition with the holistic approach. The seventeen words in legal amount are written five times as shown in Figure 4.7. There are 3,570 images of 17 classes with equally 210 samples per class. These test sets are used in the word classification process. Since these words are separated with significant space, an algorithm can be developed to extract each single word from the forms. The word images are processed to reduce noise then converted to binary format as was the previous data set. The handwriting word samples are collected from 52 volunteers. The word samples are shown in Figure 4.8.



	Word	1	2	3	4	5
1	หนึ่ง	หนึ่ง	หนึ่ง	หนึ่ง	หนึ่ง	หนึ่ง
2	สอง	สอง	สอง	สอง	สอง	สอง
3	สาม	สาม	สาม	สาม	สาม	สาม
4	สี่	สี่	สี่	สี่	สี่	สี่
5	ห้า	ห้า	ห้า	ห้า	ห้า	ห้า
6	หก	หก	หก	หก	หก	หก
7	เจ็ด	เจ็ด	เจ็ด	เจ็ด	เจ็ด	เจ็ด
8	แปด	แปด	แปด	แปด	แปด	แปด
9	เก้า	เก้า	เก้า	เก้า	เก้า	เก้า
1(--1)	เอ็ด	เอ็ด	เอ็ด	เอ็ด	เอ็ด	เอ็ด
2X	ยี่	ยี่	ยี่	ยี่	ยี่	ยี่
Ten(*10)	สิบ	สิบ	สิบ	สิบ	สิบ	สิบ
Hundred	ร้อย	ร้อย	ร้อย	ร้อย	ร้อย	ร้อย
Thousand	พัน	พัน	พัน	พัน	พัน	พัน
Ten Thousand	หมื่น	หมื่น	หมื่น	หมื่น	หมื่น	หมื่น
Hundred Thou	แสน	แสน	แสน	แสน	แสน	แสน
Million	ล้าน	ล้าน	ล้าน	ล้าน	ล้าน	ล้าน

Figure 4.7 The handwriting of amount words.



(a)

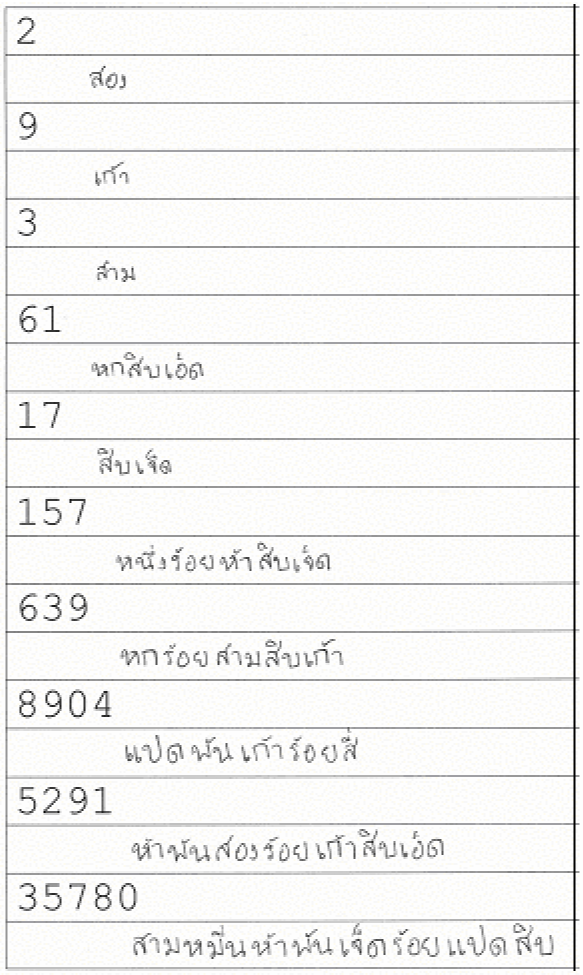
(b)

Figure 4.8 Samples of handwriting word. (a) Word สาม. (b) Word แปด.

4.6 The legal amounts data set

This section discusses the design of amount handwriting samples to test the performance of the recognition system. The total number of possible amounts from 1 to 9,999,999 is 10,000,000. However, it is impossible to collect such a large amount from a volunteer. A small set of legal amount designed with the properties covering general cases of legal amount writing is preferred. The testing amounts should contain equal distribution among all words. Moreover, they should have equal distribution in each digit length range. These properties would make the result analysis more reliable. However, if the free writing is allowed for any amounts, the statistic analysis can be complicated because the samples tend to have an unequal number of words per class; or worse, some words do not exist in their writing at all or only part of the syntax rules are used with those amounts. To prevent this problem, the data set have to be designed by the following concepts.

The legal amounts are designed with the following guidelines. The samples are the formal legal amounts of the minimum 1 digit to the maximum seven digits amount. The test set is designed for number in a range of 1 to 9,999,999. The amounts are divided into seven groups: [0-9], [10-99], [100-999], [1,000-9,999], [10,000-99,999], [100,000-999,999] and [1,000,000-9,999,999]. Two samples are randomly generated for each range. Each set of the testing data contains 16 words. There are 140 amounts generated by this guideline. The sample amounts are shown in Figure 4.9. The 140 legal amounts consisting of 986 words are collected from 10 people. The characters distributions of the test set and the word length distribution are shown in Table 4.4 and Table 4.5. The word length distributions are not equal because some digits are zero which makes some legal amounts shorter than others. In the test amounts, there are 2,489 characters consisting of 2,281 isolated and 208 connected characters. Figure 4.10 shows samples of legal amounts of five word lengths.



2

สอง

9

เก้า

3

สาม

61

หกสิบเอ็ด

17

สิบเจ็ด

157

หนึ่งร้อยห้า สิบเจ็ด

639

หกร้อย สามสิบเก้า

8904

แปด พัน เก้าร้อยสี่

5291

ห้าพันสองร้อย เก้าสิบเอ็ด

35780

สามหมื่นห้าพัน เจ็ดร้อย แปด สิบ

Figure 4.9 Sample of handwriting of the test amounts.

Word	Freq.	Word	Freq.
หนึ่ง	30	เอ็ด	14
สอง	68	ยี่	10
สาม	62	สิบ	122
สี่	51	ร้อย	100
ห้า	68	พัน	78
หก	72	หมื่น	62
เจ็ด	65	แสน	42
แปด	56	ล้าน	22
เก้า	64		

Table 4.4 Words distribution.

Length	1	2	3	4	5	6	7	8	9	10	11	12	13
Freq.	25	9	17	2	21	4	20	7	15	2	21	3	10

Table 5.5 Word length distribution.

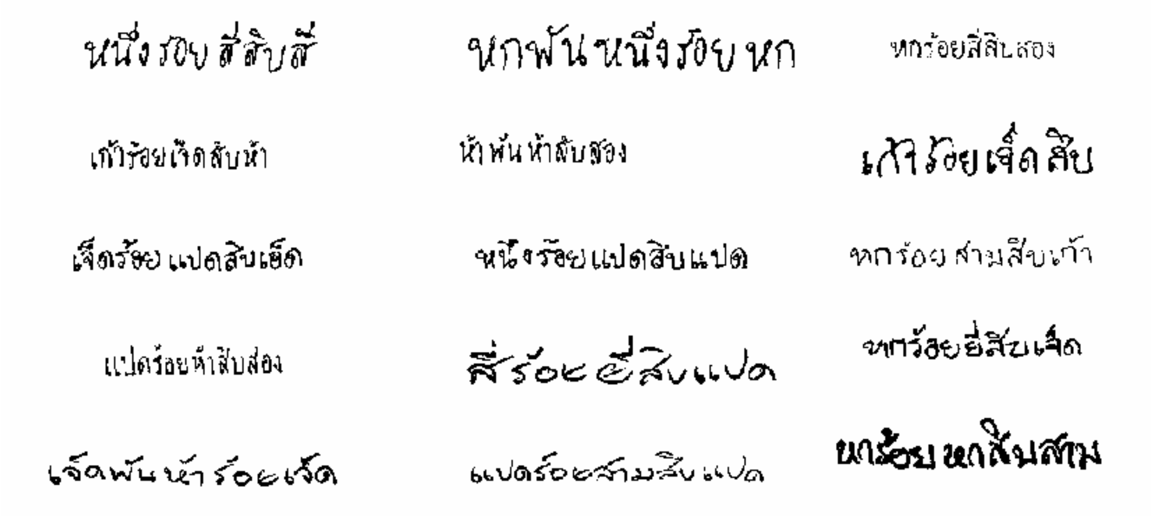


Figure 4.10 Samples of handwriting of legal amounts with five word lengths.

4.7 Thai character drawing

This section will give the basic information of how Thai characters are drawn and the necessary vocabulary of the character structures to be referenced when analyzing the problem of handwriting recognition. Basic writing guidelines show how the characters are drawn, such as the position of the beginning and the direction of the contour to the ending point and how to put the character in the reference positions in a line. In Thai handwriting guidelines, each character must be drawn separately from all others. Most Thai characters have one stroke and only a few characters have multiple strokes. Examples of one-stroke characters are [ก], [ป], [า] and [เ]. Examples of two-stroke characters are [ส], [ฐ], [ศ] and [ษ]. The beginning position of a stroke with the loop or curl shape is called the *head* and the ending stroke is called the *tail*. In amount words, [ส] and [ี] are the only two characters that need two strokes to be drawn. Figure 4.11 and Figure 4.12 show the drawing of Thai characters with one and two strokes.

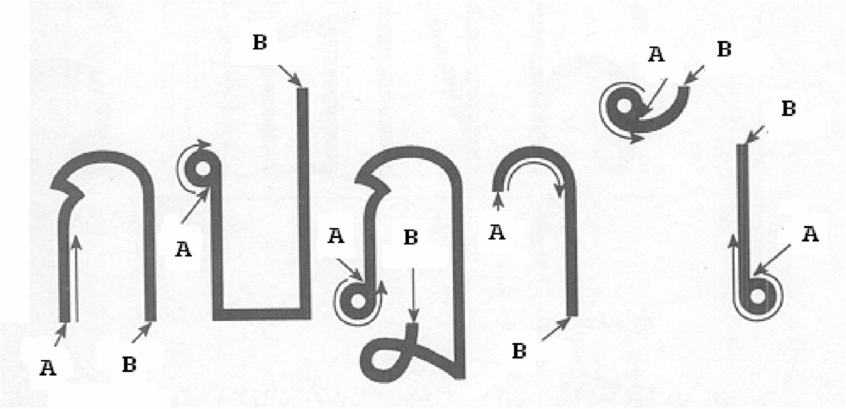


Figure 4.11 The drawing direction of one stroke characters. A is the beginning and B is the ending of the stroke. Reproduced from The Thailand Royal Institute (1999).

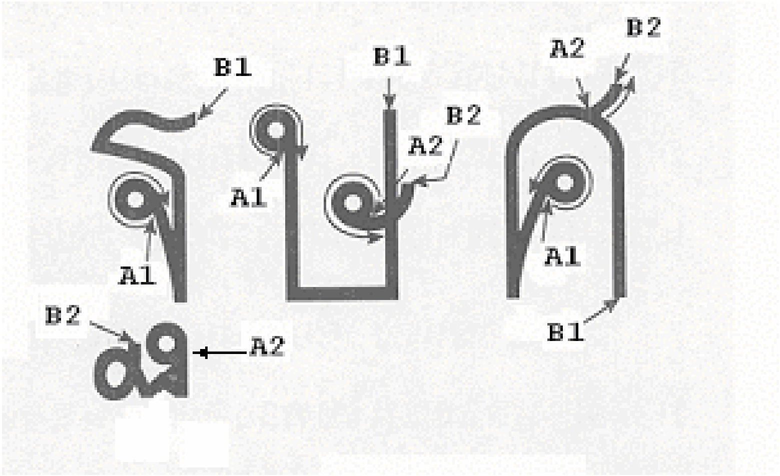


Figure 4.12 The drawing directions of two strokes characters. A1 is the beginning and B1 is the ending of the first stroke and A2 and B2 is the beginning and ending of the second stroke. Reproduced from The Thailand Royal Institute (1999).

Thai characters have the same properties shared among every written language. Each character is unique, i.e., they are designed with distinguishable features or structures from each other. Readers have little or no problems at all when reading characters compiled with standard shapes, a.k.a., printed characters. However, a few characters are alike in main structures but different in some specific features. Figure 4.13 shows distinguishable characters and similar characters.

ห น ง ส พ
ด ค ต บ ป
a)
b)

Figure 4.13 Some features in small area are used to identify the character class. (a) Some distinguishable characters. (b) Two groups of similar characters.

4.8 Problematic handwriting

The computer recognition difficulties come from the writing equipment and writing style. Writing equipment, i.e., pen type and paper qualities, produce different qualities of masks on the paper, for example, writing with dry ink usually produces broken strokes. Some image processing processes are required to improve the handwriting image before the recognition processes. Improving the handwriting image quality by image processing algorithms is a standard requirement for the recognition system in every language.

The more complicated problems result from individual writers' styles. Human peculiarities in writing are distorted from the standard shapes, for example, some extra strokes are added, the strokes length ratios are changed and some strokes are omitted. The writing style difficulties are unique for each language. Moreover, handwriting is not always clearly drawn. It might contain ambiguous characters whose shapes assemble to more than one class. Moreover, handwriting tends to have a number of touched characters that are difficult to separate into single characters. The ambiguous and cursive writing leads to segmentation error and to a reduction in recognition rate.

Many handwritten characters are indistinguishable at character levels. It is difficult to determine what class a particular character belongs to when given information of a single component without other knowledge such as position or nearby characters. When collecting data from a number of subjects, the possibility of getting ambiguous characters is very high. Characters with the same shape but which belong to different classes can confuse even humans, not to mention the classifier machines.

4.8.1 Peculiar stroke drawing

When the writing guidelines are not restricted, the characters' structures are altered by personal styles. For example, the regular character [ส] might have the beginning of *tail* drawn at the *roof* or at the *legs* part of that character as shown in Figure 4.14. These characters are drawn with two strokes. Figure 4.15 shows a variation of character ส when drawn with one stroke or with excessive strokes. Another type of irregular shape is the broken stroke drawing. This irregularity might come from writing equipment such as dry ink or the writer's ignorance. In Figure 4.15 character [ส] is drawn with a broken *tail* and character [ย] is drawn with a broken *entail.*

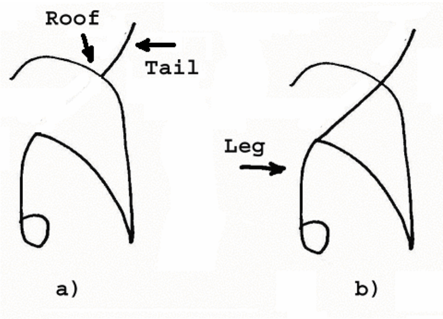


Figure 4.14 Two common ways to draw character [ส]. (a) The *tail* is connected with the *roof*. (b) The *tail* is connected to the front *leg* and run cut off the *roof*.

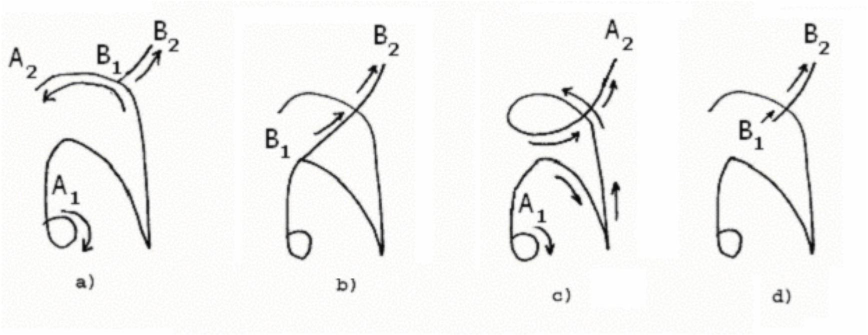


Figure 4.15 The drawing of characters [ส]. (a) and (b) The normal drawing. (c) The drawing with one stroke. (d) The drawing with excessive stroke.

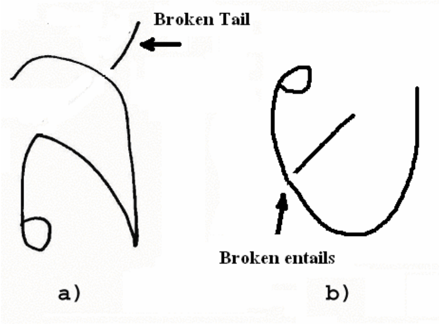


Figure 4.16 Broken stroke drawing.

4.8.2 Characters shape distortion by personal style

A character is ambiguous when its shape is distorted to the level that the main features that identify the characters' classes are not clear. This is one of the most difficult recognition problems faced when writers adapt a character's form to suit their own personal styles. This manifests itself as exaggerated strokes or very unusual character formation. These forms may become too strange to be recognized by other readers. The distortion might include only parts of a character or the whole shape. The distortion level might be considered by the need of other characters, lexicon, and syntax combined together to help identify that character. This problem requires a large number of samples to create a robust system. Figure 4.17 shows an example of ambiguous characters in the word **ร้อย.** In this Figure, three characters, [ร], [อ] and [ย], are difficult to identify, even for humans.

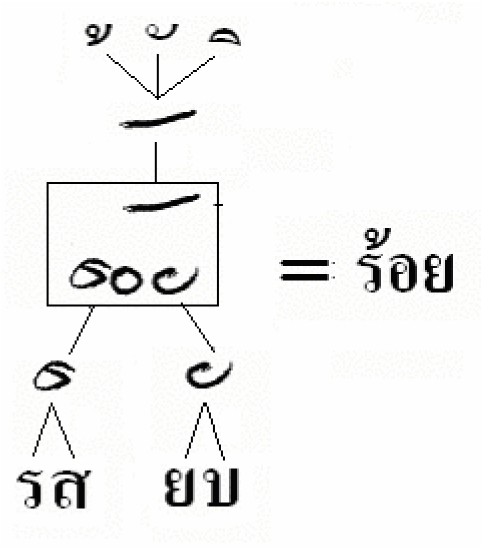
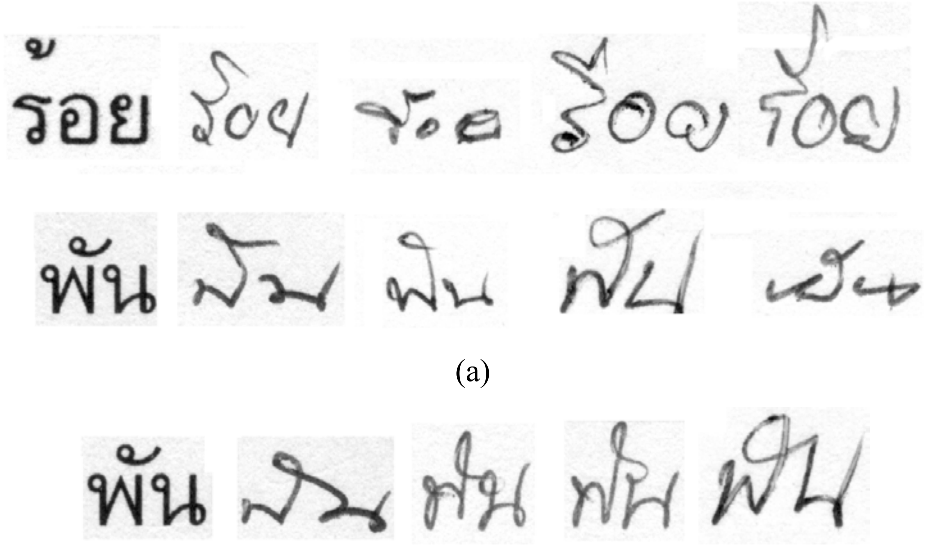


Figure 4.17 Examples of ambiguous characters.

4.8.3 Touching Characters

The writing of two or more characters that touch is very common in Thai handwriting. Touching characters tend to be within the single word rather than between words. The touched writing might be made by accident or by a writer's ignorance, because humans can read most touching characters. However, the Thai writing guidelines have no cursive style as does English. Therefore, the connected strokes of one character might be connected or cut to another character at random positions. Thai character segmentation seems to be a very challenge problem that has yet to be solved.

Another form of touching characters is the drawing of strokes without lifting the pen. These forms are often found among characters in the same word. In amount word lexicon, there are words having consonants with vowels on the upper position including **หนึ่ง**, **สี่**, **ห้า**, **เจ็ด**, **เก้า**, **สิบ**, **ร้อย**, **พัน**, **หมื่น**, and **ล้าน.** The connecting strokes are often occurred between the characters in body level and upper level. Figure 4.18 (a) shows four samples of word **ร้อย** and **พัน** words which have an excessive stroke connected between the characters in the *body* and the *upper* level. The written sequences follow the actual position order. The connected stroke sometimes connects two or more characters such as in word หมื่น. The exaggerated strokes are intentionally drawn from the end of the consonant [ร] to the vowel [้], as in word ร้อย or the stroke is drawn from the end of the consonant [พ] to the vowel [ั] , as in word พัน. In some other distorted drawing cases, more strokes also are used to connect the characters in the upper level to the body level. In general, the exaggerated strokes add no extra information about the identity of the character. These strokes can be considered as noise. Humans simply ignore this type of stroke and interpret the shape as separated characters. In the computer recognition systems, however, the detection and elimination of these excessive strokes is considered a difficult problem. Figure 4.18 (b) shows an example of the word พัน in which three characters are connected together.



(a)

(b)

Figure 4.18 Samples of connected handwriting characters. (a) The connected characters are between the consonant and vowel on above. (b) Three characters are connected as one component.

4.8.4 Multi-writer similarity problems

Recognition problems involve identifying characters with similar shapes. This problem might be solved directly by creating a new feature that identifies some unique properties between similar characters. For example, the pair character [ด] and [ต] requires features that can identify the difference at the *roof* structure, while the pair [ด] and [ค] require features that can identify the difference at the *head* structures. However, in multiple-writer environment, it is common to find that two different character classes can have the same shape. For example, character [น] and [บ] drawn by one person are distinguishable but these [น] characters are indistinguishable from [บ] characters written by another person. The other information such as word level or syntax level is required to identify the correct character class.

4.9 Features evaluation

The objective of this section is to evaluate how each feature can represent distinct characteristics of handwriting characters. Many features described in previous Chapters, such as Fourier descriptors, have been used successfully with printed characters. These features work well because the printed characters' shapes are well designed and consistent. However, the handwriting characters' shapes are altered from the standard and therefore are inconsistent even when drawn by the same person.

Each feature will be introduced, and the printed characters will be used as a simplified example to illustrate how each feature can become immune to the variant of the characters' shapes. Each feature will be used to train the character recognition system and test with the testing set. The data set in the experiment are 2,500 handwriting character samples of 25 character classes with 100 samples per class. The data is divided into the training and testing set with ratio 3:1. The training set ratio is selected to be higher than the testing set because the number of the training set is very small compared to the population size of general handwriting. The features to be evaluated are the Fourier descriptor, projection profiles with zoning, and topology features.

4.9.1 Fourier Descriptor (FD)

Shape description by Fourier Descriptor (FD) has been extensively studied for character recognition. Fourier descriptors are a series of significant coefficients related to character structures, and the information is kept in compact form. Only a small number of the FD is sufficient to describe the original characters. Usually 15-25 coefficients are selected. Variants of FD derivations discussed in this study are abbreviated as JFD (Jain 1989), EFD (Kuhl & Giardina 1982), GFD (Granlund 1972) and WFD (Wang et al 1994).

These features have been proven invariant to rotation, translation and scaling. Figure 4.19 shows rotated, scaled and translation transformations of printed characters and their JFD when the coefficient is shown from 1 to 10. The coefficients of order fourth and higher are normally very small. Therefore, their logarithmic values are used for visual observation. However, when applying FD to handwritten characters, there are a number of cases where these features are inappropriate or ineffective. The variants of handwritten characters are not simply a combination of rotation, scaling or translation. It also includes shearing, slant transformation and the altered shapes from the standard by personal styles of writing. Even the characters written by the same writer are not exactly the same.

In Figure 4.20, the first row is the rotated scaled and translation images of character [พ] and their four correspondent Fourier descriptors: JFD, MJFD, GFD and WFD respectively. The FD values are in logarithmic of 10 for visual observation. The GFD has four components but only one component is shown. The character was rotated from -25 degree to +31 degree and scaled from 0.85 to 1.25.

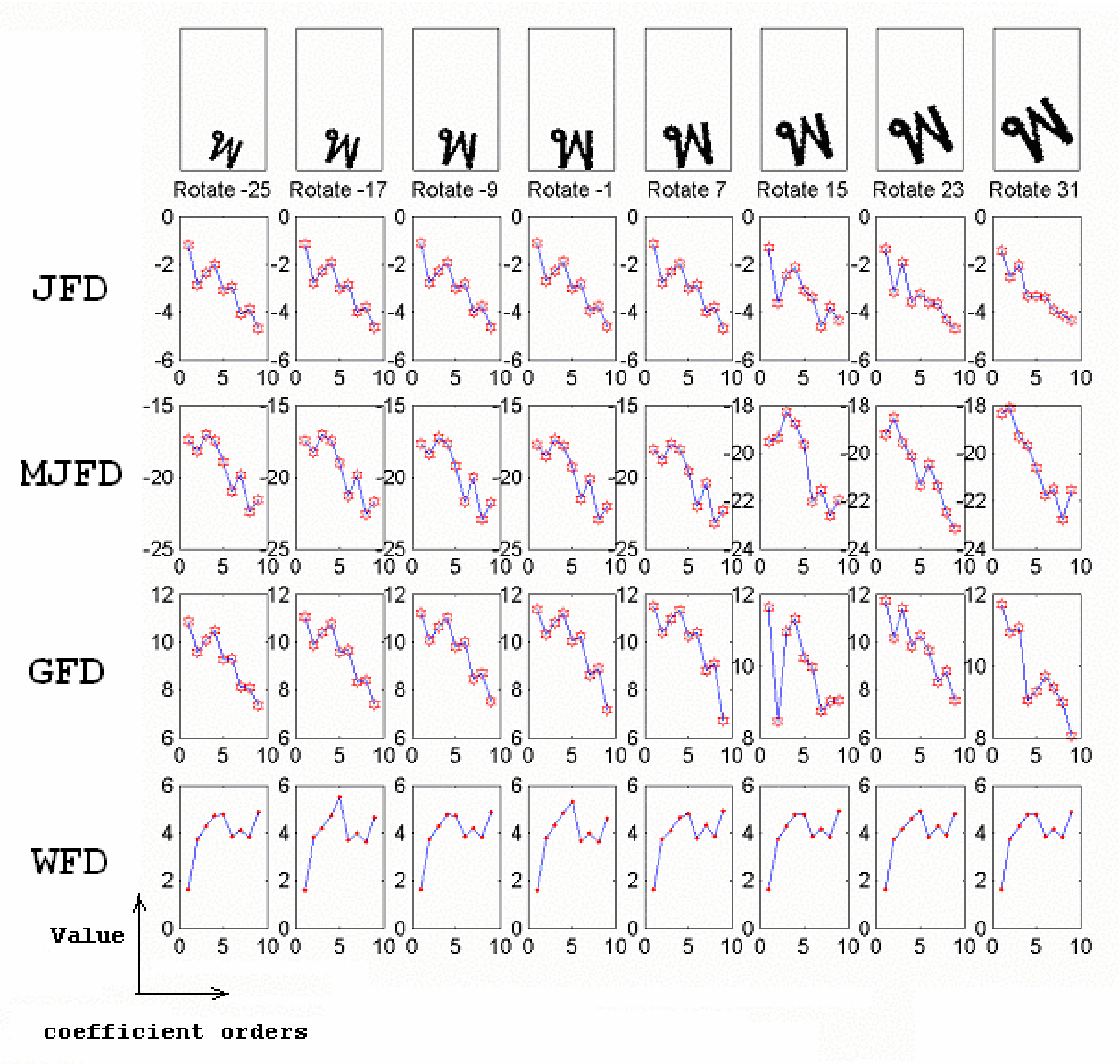


Figure 4.19 Rotation and translation transformations of character [พ] and their Fourier descriptors.

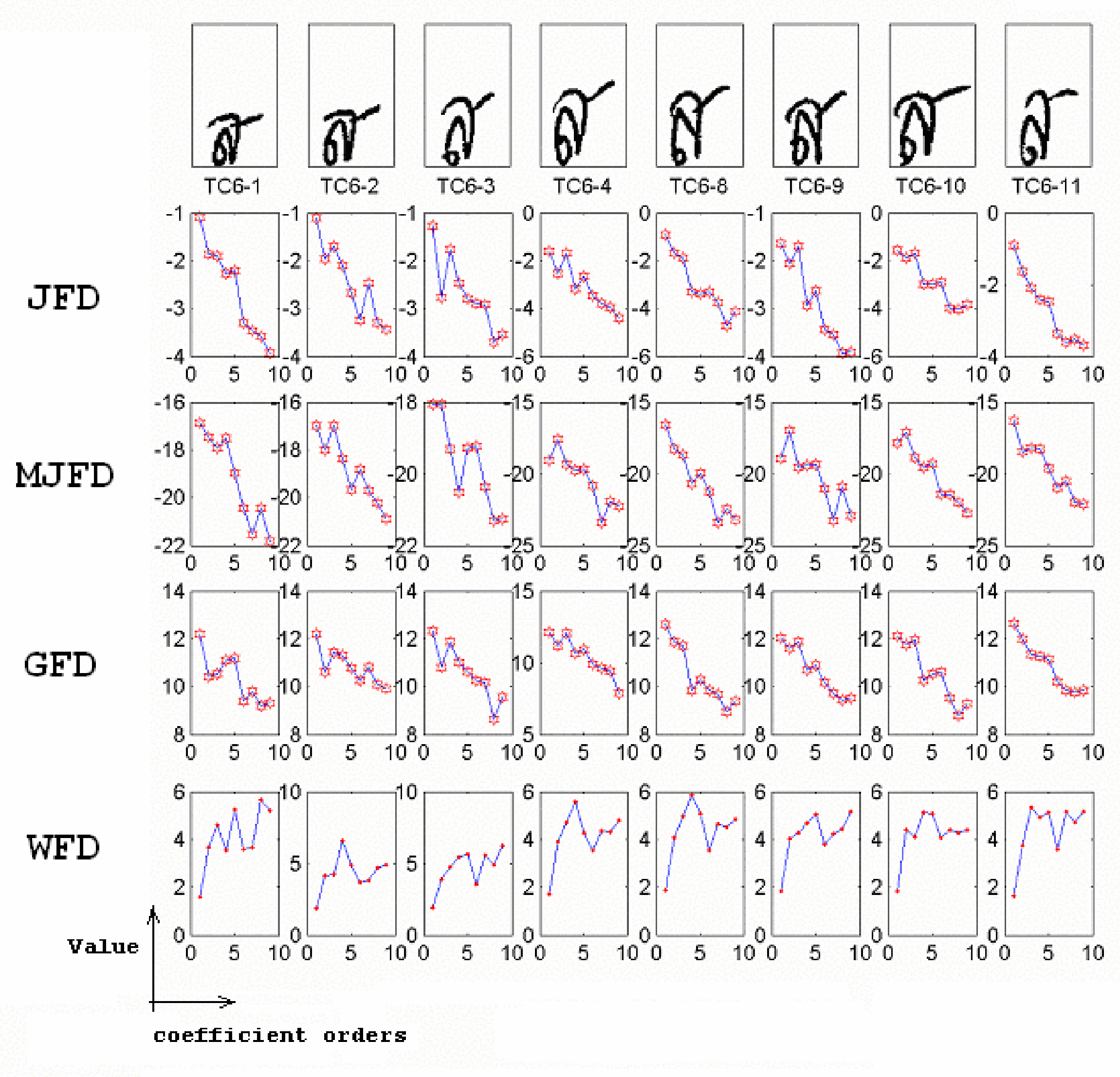


Figure 4.20 Handwritten characters [ส] and their four correspondents, JFD, MJFD, GFD and WFD. TC6-1 to TC6-2 come from one person and TC6-33 to TC6-11 come from another person.

The FD is a compact feature and invariant to basic image transformation. Only small numbers of FD coefficient contain most shape information of the original image. In the empirical experiments, 20-25 of most significant coefficients are enough to construct the original image. However, the FD features have two main weaknesses when applied to the Thai characters: a lack of specific ways to emphasize the important structures, and the ineffectiveness of internal structure or region description.

Some Thai characters share the same structure but are different in small details. These characters have similar FD coefficients, especially the lower order coefficients. There is no explicit way to specify the importance of these small details that are used to classify the character of a similar group by FD. For example, the *leg* structure can be used to identify [บ] or [น]. Even the first few FD can construct the approximation of the original images but these FD of both characters are very similar. It requires a larger number of FD to describe their difference in more detail. However, larger numbers of FD are sensitive to noise at the boundary as well.

Moreover, the invariant to rotation of the character shapes of FD appears to be a preferred property for recognizing the transformed objects. This property can be used for two dimension object recognition which might be rotated, scaled or translation. However, these properties have some drawbacks when applied to those characters sharing the same structure but different directions, for example, [ก] and [บ]. Since the FD is invariant to rotation, the values of both shapes are very similar. Other examples of Thai characters that share the same main structure are [ก], [น], [บ], and [ป]. Their FD coefficients are very similar.

The FD information cannot describe the internal structures or regions. Most FD coefficients are derived from a character's outer contour boundary. Obviously, the information within the outer contour is neglected. For example, FD is not concerned with the holes inside the structures. For Thai character, these holes are an indicator of the *head* position and can be used to identify specific character classes. Moreover, some irregular strokes can hide some structures or alter the outer contours. Characters might be written in such a way that internal structures are hidden from the outer contour.

Figure 4.21 shows examples of variations of ส. The first ส is shaped normally but the others are altered by some broken and excessive strokes. The loop-like structure is called the *head.* If this loop is not completed it would be called the *broken head*. If the *tail* structure, i.e., a line joined in the upper part of a character, does not attach to the main structure, it would be called the *broken tail*. The *broken tail* alters the structure more than the *broken head* and excessive strokes because part of the structure is separated from the whole shape during the character segmentation.

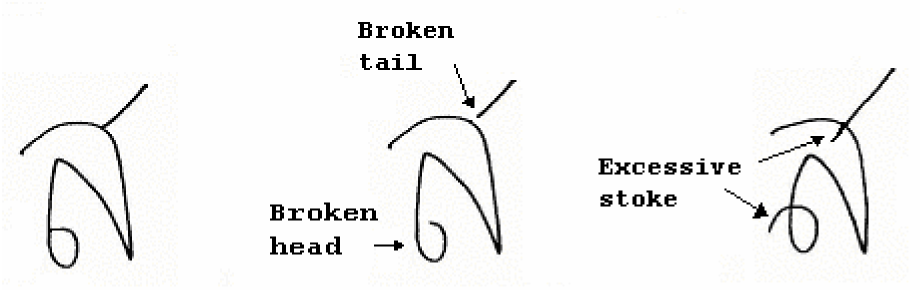


Figure 4.21 Variation of character [ส] with some peculiar strokes.

Figure 4.22 shows variations of character [ส] and their outer contours. On the first row, the normal shape is shown as the first characters. The second character is drawn with the *broken head* and the *broken tail*. The third character is drawn with an excessive line at the *head* position. The fourth character is drawn by a single stroke with a loop to make a *tail* which hides the internal structure. The fifth character is drawn with the *tail* touching the *leg* and it hides the internal structure from the outer boundary. On the second row are the corresponding outer boundary images. Obviously, their Fourier descriptors calculated from outer boundaries are different.

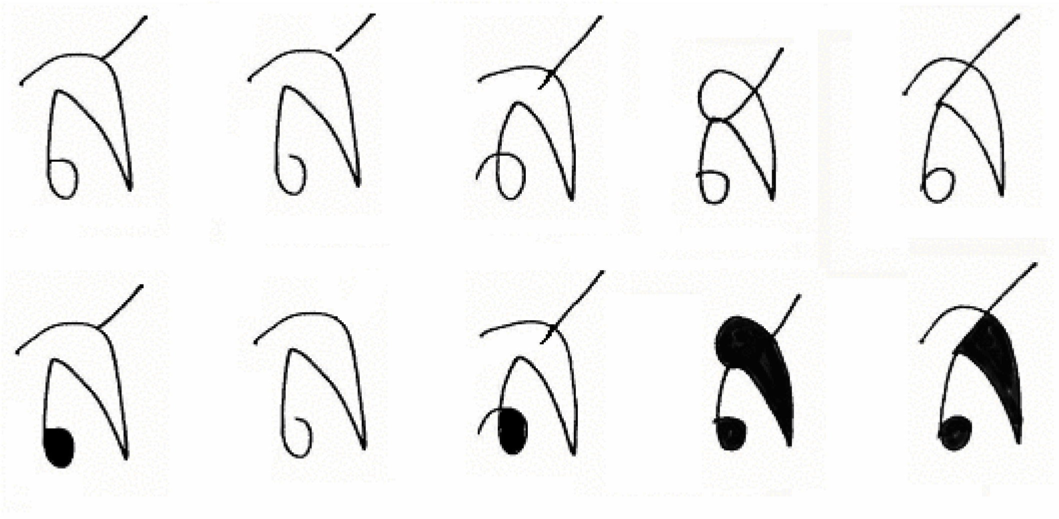


Figure 4.22 Variations of characters [ส] and their outer contours.

The FD is derived from a closed curve, so this feature is inapplicable to those characters having multiple-components. Many handwriting styles have multiple-component characters. The multiple-components are a result of personal styles or from broken strokes created while writing. Three types of Fourier descriptor generate one set of Fourier coefficients while the elliptic Fourier descriptor generates four set Fourier coefficients for a closed curve. The term 1 to 30 of the Fourier coefficients are selected to represent each character. The Fourier coefficients are extracted from 1,875 handwriting training data set and used to train the neural network classifier systems. The networks have the number of input nodes equal to the number of features, 20 hidden nodes and 25 output nodes. The classifiers are tested with the 625 handwriting characters from the testing data set. Each feature is used to train the classifiers with 0.005 error rate and maximum 1000 epochs. The best result among five experiments is selected and shown in Table 4.6 through Table 4.9 in form of confusion matrix. The number in the table is the average recognition rate in percent. The rows represent the true classes and the columns represent the predicted classes.

	ห	น	ึ	่	ง	ส	อ	า	ม	ี	้	ก	เ	จ	็	ด	ป	ิ	บ	ย	ร	พ	ั	ี	ล
ห	20	16	-	-	-	4	-	-	40	-	-	4	-	-	-	-	-	-	-	-	16	-	-	-	-
น	12	40	-	-	-	8	-	-	16	-	4	-	-	-	-	-	-	-	-	4	-	8	-	4	4
ึ	4	-	48	8	4	-	-	4	-	4	-	-	8	4	4	-	-	-	-	4	4	-	-	4	-
่	-	-	-	64	-	-	8	4	4	4	-	-	4	-	-	4	-	8	-	-	-	-	-	-	-
ง	-	-	-	-	20	4	-	-	-	4	-	4	4	12	-	24	4	-	-	-	8	-	-	-	16
ส	16	8	-	-	-	32	-	-	12	-	-	-	-	-	-	-	-	-	4	-	12	12	-	-	4
อ	-	-	-	16	-	-	32	-	-	-	8	4	-	-	4	8	-	12	-	-	-	8	-	-	8
า	4	-	4	4	-	-	-	40	-	-	20	4	4	4	-	8	-	4	-	-	-	-	-	-	4
ม	16	20	4	-	-	4	-	-	40	-	-	4	-	-	-	-	-	-	-	-	4	4	-	-	4
ี	8	4	12	-	-	4	-	-	4	24	4	-	4	4	-	8	-	-	-	-	8	-	-	-	16
้	8	-	-	-	4	4	4	12	-	4	20	-	-	-	-	4	-	-	20	-	8	4	4	-	4
ก	4	-	-	-	4	-	-	4	-	-	-	48	-	-	4	-	8	-	12	8	4	-	-	-	4
เ	-	-	4	24	4	-	-	4	4	8	4	4	32	-	4	-	-	-	-	-	-	-	4	-	4
จ	4	4	-	-	-	8	-	12	12	-	12	4	-	8	-	-	-	4	4	-	4	-	16	-	8
็	4	8	-	4	-	4	-	4	8	-	4	4	-	-	16	-	8	-	-	8	12	-	4	8	4
ด	-	-	-	-	8	-	-	-	-	8	4	12	-	-	-	44	12	-	-	-	-	-	4	-	8
ป	-	4	-	-	-	-	-	4	-	-	4	-	-	-	16	12	60	-	-	-	-	-	-	-	-
ิ	-	-	-	8	4	4	4	4	4	-	4	4	4	8	-	8	-	24	-	-	-	-	12	-	8
บ	-	4	-	-	-	4	-	-	16	-	4	12	-	4	4	8	4	-	36	-	-	-	-	-	4
ย	-	4	-	4	4	4	-	-	12	-	4	-	-	8	-	4	-	-	-	52	-	-	4	-	-
ร	4	4	-	-	4	-	-	-	8	4	4	-	-	-	-	-	-	-	-	-	60	-	-	4	8
พ	28	8	-	-	-	12	-	-	12	-	4	-	-	-	-	-	-	-	-	-	-	32	-	-	4
ั	-	-	-	-	-	-	-	12	8	-	8	-	4	20	-	12	-	-	-	-	-	-	28	4	4
ี	4	-	-	-	-	-	-	-	4	4	8	-	-	-	-	-	4	-	-	-	-	-	8	68	-
ล	12	4	4	-	-	8	-	8	24	-	4	-	-	4	-	-	-	-	8	-	4	-	-	-	20

Table 4.6 Confusion matrix for the classifier system using the Fourier descriptor type I. The average recognition rate is 36.32 percent.

	ห	น	ึ	่	ง	ส	อ	า	ม	ี	้	ก	เ	จ	็	ด	ป	ิ	บ	ย	ร	พ	ั	ื	ล
ห	20	4	-	-	4	4	-	-	16	-	-	-	-	-	4	-	-	-	-	4	8	32	-	-	4
น	8	28	-	-	-	-	-	-	8	-	4	-	-	4	-	-	-	-	-	-	-	40	-	-	8
ึ	-	-	32	12	4	-	-	-	-	12	-	-	8	-	4	8	-	-	-	-	-	4	12	4	-
่	-	-	-	56	4	-	4	-	-	-	-	-	16	-	-	8	-	-	-	-	-	4	8	-	-
ง	-	-	-	-	12	-	-	12	-	4	12	8	4	8	-	16	-	4	-	-	8	8	4	-	-
ส	-	12	4	-	-	32	-	-	-	-	4	-	-	4	4	12	4	-	-	4	4	12	-	-	4
อ	-	4	-	4	4	-	32	-	-	-	-	-	-	-	12	20	-	12	-	4	-	-	4	4	-
า	-	-	-	4	-	-	-	60	-	-	8	-	8	-	-	4	-	4	-	-	4	-	8	-	-
ม	-	8	-	-	-	-	-	-	44	-	4	-	-	8	-	4	-	-	-	-	8	20	-	-	4
ี	4	-	8	-	-	4	4	4	-	36	-	-	4	-	4	8	-	-	4	4	4	4	-	4	4
้	-	-	-	-	-	4	-	16	4	8	20	4	8	-	-	4	-	-	8	-	8	4	12	-	-
ก	-	-	-	-	4	4	-	4	-	8	-	48	-	-	-	12	-	-	12	-	-	4	4	-	-
เ	-	-	8	28	4	-	-	8	-	-	4	-	24	-	-	4	-	-	-	-	4	8	8	-	-
จ	-	4	-	-	-	-	4	12	-	12	-	-	4	24	-	8	-	4	-	8	-	4	12	-	4
็	-	4	-	-	-	8	-	8	4	-	8	-	-	-	28	8	4	4	4	8	-	-	4	4	4
ด	-	-	-	-	4	-	-	-	-	-	-	4	-	4	-	72	-	-	4	8	-	4	-	-	-
ป	-	-	-	-	4	4	-	8	-	-	-	-	-	-	4	12	60	-	-	-	-	-	4	-	4
ิ	-	-	4	8	-	-	8	4	-	4	-	12	-	-	-	4	-	28	8	4	-	8	8	-	-
บ	-	12	-	-	8	8	-	-	-	-	4	8	-	4	12	8	4	-	24	-	-	4	-	-	4
ย	-	4	-	-	4	4	-	4	4	-	-	-	4	4	-	-	-	-	-	52	-	8	-	-	12
ร	-	-	4	-	12	-	-	-	4	4	12	-	-	-	-	4	-	-	-	-	48	4	4	-	4
พ	8	4	-	-	-	4	-	-	12	4	-	-	-	-	-	4	-	-	-	-	-	60	4	-	-
ั	-	-	4	-	-	-	-	8	-	4	12	-	4	-	-	4	-	4	-	-	4	4	52	-	-
ื	-	-	-	-	-	4	4	-	-	4	4	-	-	-	4	4	-	-	-	-	8	12	-	56	-
ล	-	-	4	-	4	-	-	-	4	4	-	-	-	-	-	12	4	-	4	-	-	-	4	-	60

Table 4.7 Confusion matrix for the classifier system using the Fourier descriptor type II. The average recognition rate is 40.32 percent.

	ห	น	ึ	่	ง	ส	อ	า	ม	ี	้	ก	เ	จ	็	ด	ป	ิ	บ	ย	ร	พ	ั	ื	ล
ห	36	16	-	-	4	8	-	-	4	-	-	-	-	4	-	-	-	8	-	-	-	16	-	4	-
น	4	36	-	-	-	8	-	-	8	4	-	4	-	8	4	-	4	4	-	-	4	8	-	-	4
ึ	-	-	40	4	-	8	-	-	-	-	-	-	4	4	-	4	-	8	-	-	20	-	8	-	-
่	-	-	-	84	-	-	4	-	-	-	-	-	4	-	-	-	-	4	4	-	-	-	-	-	-
ง	-	-	-	-	36	-	-	12	-	-	8	-	8	4	-	8	-	4	4	4	12	-	-	-	-
ส	8	-	-	-	-	40	-	-	-	4	-	-	-	-	8	4	-	12	-	-	8	12	-	4	-
อ	4	-	-	8	4	4	24	-	-	-	-	-	4	4	8	16	-	16	-	-	4	-	-	4	-
า	-	-	-	8	4	-	-	68	-	-	-	-	4	-	-	-	8	4	-	-	4	-	-	-	-
ม	24	16	-	-	4	8	-	-	12	4	-	-	-	4	-	4	4	-	-	-	4	8	-	4	4
ี	4	-	4	-	4	-	4	-	4	20	8	4	4	-	-	4	-	12	-	-	24	-	-	4	-
้	8	-	-	-	8	-	-	24	-	4	16	-	8	-	4	4	4	-	-	-	8	4	8	-	-
ก	-	4	-	-	-	-	-	4	-	-	4	64	-	-	-	4	8	-	8	4	-	-	-	-	-
เ	-	-	-	24	8	-	4	12	-	-	-	-	48	-	-	-	-	4	-	-	-	-	-	-	-
จ	-	-	-	-	-	-	-	24	-	-	-	4	8	20	-	8	4	16	8	-	-	-	4	-	4
็	-	4	-	-	4	16	-	8	-	-	-	-	8	-	16	4	12	-	-	12	4	4	-	4	4
ด	-	-	-	-	12	-	-	-	-	4	-	12	-	-	-	36	4	8	4	4	8	4	-	-	4
ป	-	-	-	-	-	-	-	20	-	-	-	4	-	-	-	-	68	4	-	-	-	-	-	-	4
ิ	-	-	-	8	-	-	-	20	-	-	8	4	12	4	-	8	8	8	-	-	4	-	8	4	4
บ	-	8	-	-	4	-	-	-	-	-	4	12	-	4	4	-	8	12	32	-	-	-	-	4	8
ย	-	8	-	-	-	-	-	4	-	-	-	8	4	4	-	8	-	4	4	36	-	4	-	8	8
ร	8	-	8	-	4	-	-	-	-	4	-	-	8	4	-	-	-	-	-	-	56	-	4	4	-
พ	-	8	-	-	-	8	-	-	20	-	-	-	-	-	-	-	-	-	-	8	4	52	-	-	-
ั	-	-	4	-	-	-	-	12	-	-	4	-	8	-	-	4	-	12	-	-	-	-	52	4	-
ื	8	4	-	-	-	-	-	-	-	12	-	-	-	4	-	-	4	8	-	4	4	4	-	40	8
ล	4	-	4	-	4	4	-	-	4	-	4	4	-	4	4	8	4	4	4	-	4	4	-	4	32

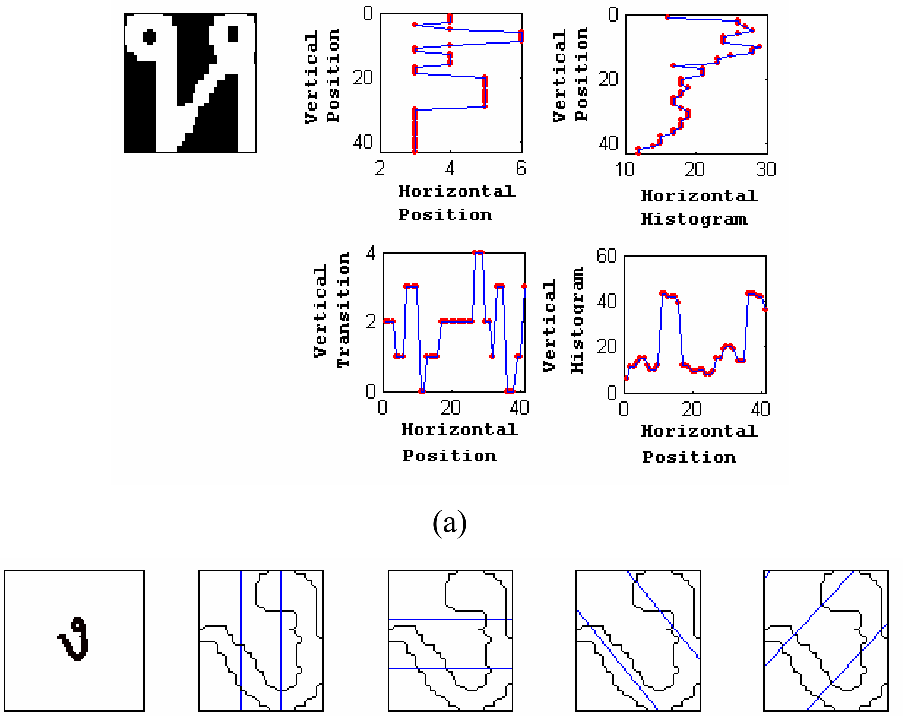
Table 4.8 Confusion matrix for the classifier system using the Fourier descriptor type III. The average recognition rate is 38.88 percent.

	ห	น	ึ	่	ง	ส	อ	า	ม	ี	้	ก	เ	จ	็	ด	ป	ิ	บ	ย	ร	พ	ั	ื	ล
ห	28	12	8	-	4	-	-	-	8	-	-	-	4	8	4	4	-	4	8	-	-	4	4	-	-
น	12	28	4	-	-	-	4	-	4	8	12	-	-	-	-	-	-	-	4	-	-	16	-	4	4
ึ	8	-	56	-	-	-	-	4	4	8	-	-	4	-	-	-	-	12	-	-	-	-	4	-	-
่	-	-	-	52	-	-	4	-	-	-	-	-	16	12	8	-	4	4	-	-	-	-	-	-	-
ง	-	4	8	4	12	-	8	4	-	-	16	4	4	4	8	-	12	4	4	4	-	-	-	-	-
ส	4	-	8	8	4	12	16	4	-	8	-	-	-	-	4	4	4	4	-	-	4	16	-	-	-
อ	4	-	-	8	-	4	36	8	-	-	-	8	-	-	4	-	-	4	4	4	4	4	4	-	4
า	-	-	4	-	-	-	-	68	4	-	-	4	8	-	-	-	-	8	-	-	-	4	-	-	-
ม	-	-	8	-	-	-	4	-	48	4	-	-	-	-	8	-	-	-	12	4	-	4	4	-	4
ี	-	-	24	-	-	-	4	-	-	44	4	-	-	4	4	-	-	-	-	4	-	-	-	12	-
้	-	4	8	-	4	-	4	-	12	8	16	-	4	-	-	-	4	8	4	4	-	4	16	-	-
ก	-	-	-	-	-	4	-	8	-	-	-	72	-	-	8	-	-	-	-	-	4	-	-	-	4
เ	-	-	4	24	-	-	4	24	-	12	-	-	12	4	-	-	-	12	-	-	4	-	-	-	-
จ	-	-	4	8	4	-	20	8	-	4	-	-	4	16	8	4	-	-	-	4	8	-	-	-	8
็	4	-	4	8	4	-	12	8	-	-	4	4	-	8	28	4	4	-	-	-	-	-	4	4	-
ด	-	-	-	-	-	-	12	-	-	-	4	12	8	12	20	16	-	-	-	4	4	4	-	-	4
ป	4	4	-	-	4	-	-	4	8	4	8	-	-	16	-	-	36	-	-	8	-	-	4	-	-
ิ	-	-	8	4	4	4	20	-	-	4	-	-	4	4	-	-	4	32	-	-	8	-	-	-	4
บ	-	8	4	-	-	-	-	-	-	8	4	-	-	-	4	-	-	-	40	8	-	4	4	8	8
ย	8	4	-	-	4	-	4	4	4	16	-	-	4	4	4	4	-	-	12	20	-	-	-	8	-
ร	-	-	-	4	-	-	-	24	-	4	-	8	-	16	-	-	-	4	-	-	32	-	-	4	4
พ	12	-	-	-	-	-	-	-	4	8	-	-	-	-	-	-	-	-	4	-	-	64	4	-	4
ั	-	-	12	4	4	-	4	-	4	4	12	-	4	-	-	-	-	-	4	4	-	-	44	-	-
ื	8	-	4	-	-	-	-	4	-	20	-	-	-	-	-	-	4	-	4	-	-	-	-	56	-
ล	8	-	4	12	-	-	8	4	4	8	-	4	-	-	4	-	-	-	8	4	4	4	4	-	20

Table 4.9 Confusion matrix for the classifier system using the Fourier descriptor type IV. The average recognition rate is 35.52 percent.

4.9.2 Projection profile

Projection profile is a transformation of a two-dimension binary image into two one-dimension vectors. The project profiles are not immune to image transformation. The preprocessing process is required to normalize the image to a preference size and direction before feature extraction. These features have to be used with the zoning to be immune to image transformation. A large set of labeled samples with variations are commonly needed to represent the general data model. Figure 4.23 shows two types of projection profiles: histogram and transition profiles.



(a)

(b)

Figure 4.23 (a) Histogram and transition profiles of character [ฟ]. (b) The zoning areas illustrated using character [ง] template.

Projection profiles can be categorized into histogram and transition profiles. A histogram profile is the summation of pixels along the projection axis while the transition profile is the number of transition of black and white pixels along the projection axis. The projection axis can be as simple as horizontal and vertical or extended to both diagonal axes. Both histogram and transition profiles are very easy to implement and capture the main features of characters. Projection profiles are insensitive to noise from the images boundary when the size of a character is large enough. However, they are sensitive to shearing and rotation transformation. More specifically, the *head*, one of the most important features, is undetected by histogram profiles. However, the main disadvantage is that they are not applicable with similar characters. These features are always coupled with zoning to enhance some information from specific areas which contain certain features to identify a character.

To reduce the effect from these variants, the image size is normalized either by a total number of pixels or by other methods. Both anti-skewed and anti-rotation processes are normally needed to transform the image into some preference positions. However, these algorithms do not work well with Thai characters having an unsymmetrical shape such as [ง] or [จ]. Therefore, the results are still unreliable. This study depends on statistical properties collected from a large set of samples, instead of relying on normalization processes. The histogram and transition profiles feature are applied to the same training and testing data set to compare the results. The neural network classifiers have input nodes with numbers equal to the input features, 20 hidden nodes and 25 output nodes. Five classifiers were trained with 0.005 error rate with 1000 maximum epochs. The best result from these classifiers is selected and shown in Table 4.10 and Table 4.11 in confusion matrix form.

	ห	น	ึ	่	ง	ส	อ	า	ม	ี	้	ก	เ	จ	็	ด	ป	ิ	บ	ย	ร	พ	ั	ื	ล
ห	44	24	-	-	-	-	-	-	-	-	-	8	-	-	-	16	-	-	-	-	-	4	4	-	-
น	8	64	-	-	-	-	-	-	4	-	4	-	-	-	-	-	-	-	12	-	-	8	-	-	-
ึ	-	4	60	-	-	-	-	-	-	8	-	-	-	-	-	-	-	8	-	4	-	4	8	-	4
่	-	-	-	56	-	8	-	-	-	4	8	-	4	8	4	-	-	4	-	-	-	-	-	-	4
ง	-	-	-	4	60	4	4	-	-	-	4	-	-	-	8	-	-	-	-	-	12	-	-	4	-
ส	4	4	-	8	-	64	-	-	-	-	-	-	-	-	4	-	8	-	-	-	-	-	-	-	8
อ	4	-	4	4	-	-	72	-	-	-	-	-	-	4	-	-	-	8	-	-	-	-	-	-	4
า	-	-	-	4	4	4	-	76	-	-	-	4	-	-	-	-	-	-	-	-	4	-	-	-	4
ม	8	4	-	-	4	-	-	-	52	-	4	-	-	-	-	-	-	-	8	4	-	12	4	-	-
ี	-	8	4	-	4	-	-	-	-	44	8	-	-	4	-	4	-	4	-	4	-	-	-	16	-
้	-	4	4	-	-	-	-	-	4	-	64	-	4	-	-	-	4	-	-	-	-	-	16	-	-
ก	-	-	-	-	-	-	-	-	-	-	-	96	-	-	-	4	-	-	-	-	-	-	-	-	-
เ	-	8	-	8	-	8	4	4	-	-	8	-	44	-	4	-	-	-	-	-	4	4	4	-	-
จ	-	-	-	-	4	-	4	12	-	-	-	-	-	56	-	4	-	12	-	-	-	-	4	4	-
็	4	-	-	-	8	12	-	-	-	-	-	-	-	-	60	-	4	-	-	4	-	-	8	-	-
ด	-	-	-	-	4	-	4	-	-	-	-	-	-	-	-	84	-	-	-	-	-	-	-	-	8
ป	-	8	-	-	4	-	-	-	-	4	8	-	-	-	-	-	72	-	-	-	-	4	-	-	-
ิ	-	-	8	8	4	-	4	12	-	-	-	4	-	-	-	8	-	36	-	-	-	4	-	8	4
บ	4	16	-	-	-	-	-	-	20	-	-	-	-	-	-	-	-	-	48	4	-	-	8	-	-
ย	-	4	4	-	-	-	4	-	4	-	-	-	12	-	-	-	-	-	12	48	-	-	12	-	-
ร	-	-	-	-	4	4	-	4	-	-	-	-	4	-	8	-	-	-	-	-	76	-	-	-	-
พ	4	8	-	-	-	-	-	-	8	-	-	4	-	-	-	-	4	-	-	-	-	72	-	-	-
ั	4	-	-	-	4	-	-	-	-	-	8	-	-	-	-	-	-	-	4	-	-	4	76	-	-
ื	-	-	-	-	-	-	4	-	-	-	-	4	-	-	-	-	-	4	-	12	-	-	-	76	-
ล	-	-	-	-	8	-	-	-	-	-	4	8	-	-	-	-	-	-	-	-	4	-	-	-	76

Table 4.10 Confusion matrix for the classifier system using the outer contour. The average recognition rate is 63.04 percent.

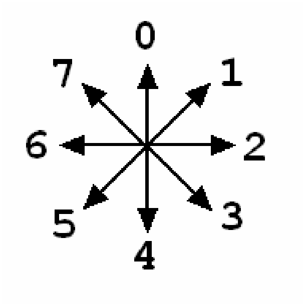
	ห	น	ื	่	ง	ส	อ	า	ม	ี	้	ก	เ	จ	็	ด	ป	ิ	บ	ย	ร	พ	ั	ื	ล
ห	52	12	-	-	4	-	-	-	-	-	4	-	-	-	-	-	4	-	-	-	-	20	-	4	-
น	16	36	-	-	-	-	-	-	8	-	8	-	-	-	-	-	8	-	12	-	-	-	8	4	-
ื	4	-	64	-	-	-	-	-	-	8	-	-	-	-	8	-	-	8	-	-	-	-	-	8	-
่	-	-	-	72	-	-	-	12	-	-	-	-	12	-	4	-	-	-	-	-	-	-	-	-	-
ง	-	-	-	-	60	4	4	8	-	-	-	-	-	-	4	-	-	-	-	-	20	-	-	-	-
ส	4	-	-	-	-	60	-	-	-	-	-	-	-	-	16	-	-	4	-	4	4	-	4	-	4
อ	-	-	-	-	-	4	64	4	-	-	-	4	-	4	-	-	-	-	-	-	8	-	-	-	12
า	-	-	-	4	-	-	-	80	-	-	-	4	-	-	4	-	-	-	-	-	8	-	-	-	-
ม	12	4	-	-	-	-	-	-	64	4	-	-	-	-	-	4	-	-	8	-	-	-	-	4	-
ี	4	-	20	-	-	-	-	-	-	40	8	-	-	4	4	4	-	-	-	-	-	4	-	12	-
้	4	4	4	-	-	-	-	-	-	-	48	-	-	-	4	-	4	4	8	-	-	8	12	-	-
ก	8	-	-	-	-	-	-	4	-	-	-	80	-	-	-	-	-	-	-	-	-	-	4	-	4
เ	-	-	-	8	4	-	8	12	4	-	-	-	52	-	4	-	-	4	-	-	-	-	4	-	-
จ	-	-	4	-	-	4	-	8	-	4	-	-	-	64	-	4	-	8	-	-	4	-	-	-	-
็	-	-	4	-	-	-	-	-	-	-	12	-	-	-	60	8	4	4	-	4	-	-	-	-	4
ด	4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	88	-	-	-	-	-	-	-	-	8
ป	-	-	-	-	-	-	-	-	-	4	8	-	-	-	4	-	80	-	4	-	-	-	-	-	-
ิ	-	-	-	12	-	4	8	-	-	4	-	-	-	12	4	4	-	36	-	-	-	-	4	-	12
บ	-	4	-	-	-	-	4	-	4	-	4	4	-	-	-	-	-	-	76	-	-	-	-	4	-
ย	4	-	4	-	-	-	12	-	-	-	4	-	-	-	-	-	-	4	16	52	4	-	-	-	-
ร	-	-	-	-	8	4	-	8	-	-	-	-	4	-	-	-	-	4	-	-	68	-	4	-	-
พ	16	-	-	-	4	-	-	-	4	4	-	-	-	-	-	-	-	-	4	-	-	68	-	-	-
ั	-	-	-	-	4	-	-	-	-	-	8	-	4	-	-	-	-	-	4	-	4	-	76	-	-
ื	4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4	-	4	-	88	-
ล	4	4	12	-	-	4	4	-	-	-	-	-	-	-	-	4	-	4	-	4	-	-	-	16	44

Table 4.11 Confusion matrix for the classifier system using the transition profiles. The average recognition rate is 62.88 percent.

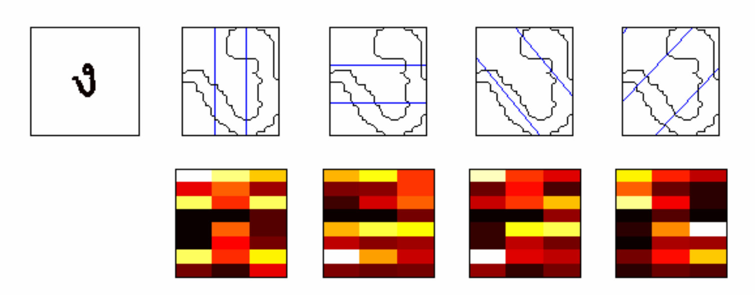
4.9.3 The edge direction feature

Information from character image indicates the direction of strokes. It is a simple task for a trained person to point out the possible beginning and ending of strokes of the finished drawing characters. However, it is a very difficult task to emulate this ability in a computer. There are attempts to reduce the information from stroke drawing into a smaller set, such as skeleton strokes which use one pixel width stroke then calculate the possible directions and sequences of those strokes. The problem is that the skeleton algorithm always creates inconsistence structures when applied to handwritten data.

Another approach is to use information from the contour points. The directions of these contour points are independent of the strokes width. Two types of edge directions are employed in this study. In the first type, the character images must be normalized into a specific size, and then a 3 by 3 template is used to scan and count the matching pattern with 4 directional element patterns: vertical, horizontal, forward diagonal and backward diagonal. The second type uses the approximation of the character image by spline approximation. The direction of the high curvature points are quantized into 8-direction chain coded and the zoning technique is used to find the proportion of each direction in each zone. Figure 4.24 shows the printed character [ง] and the edge direction features as shown in 8x3 matrixes for each zone. The intensity in each matrix cell represents the proportion of each edge direction in each zone.



(a)



(b)

Figure 4.24 (a). The 8 chain-code directions (b). The zoning template on a printed [ง] character and the corresponding edge direction features of the character.

Figure 4.25 shows the variation of printed characters [ส] and [ด]. The characters were transformed by 3 to 24 degree rotating transformation and 1.15 to 2.05 scaling transformation from the original image. As shown by the intensity images, the zoned edge-direction features are medium immune to rotation and have the potential to be used for printed character recognition. The results of the edge-direction are 48 elements feature vector with components having values normalized in range 0 to 1. Figure 4.26 shows the edge direction feature of every character. Table 4.12 and Table 4.13 show the recognition result using the edge direction type I and II features. The experiment setups are the same as the previous feature evaluation.

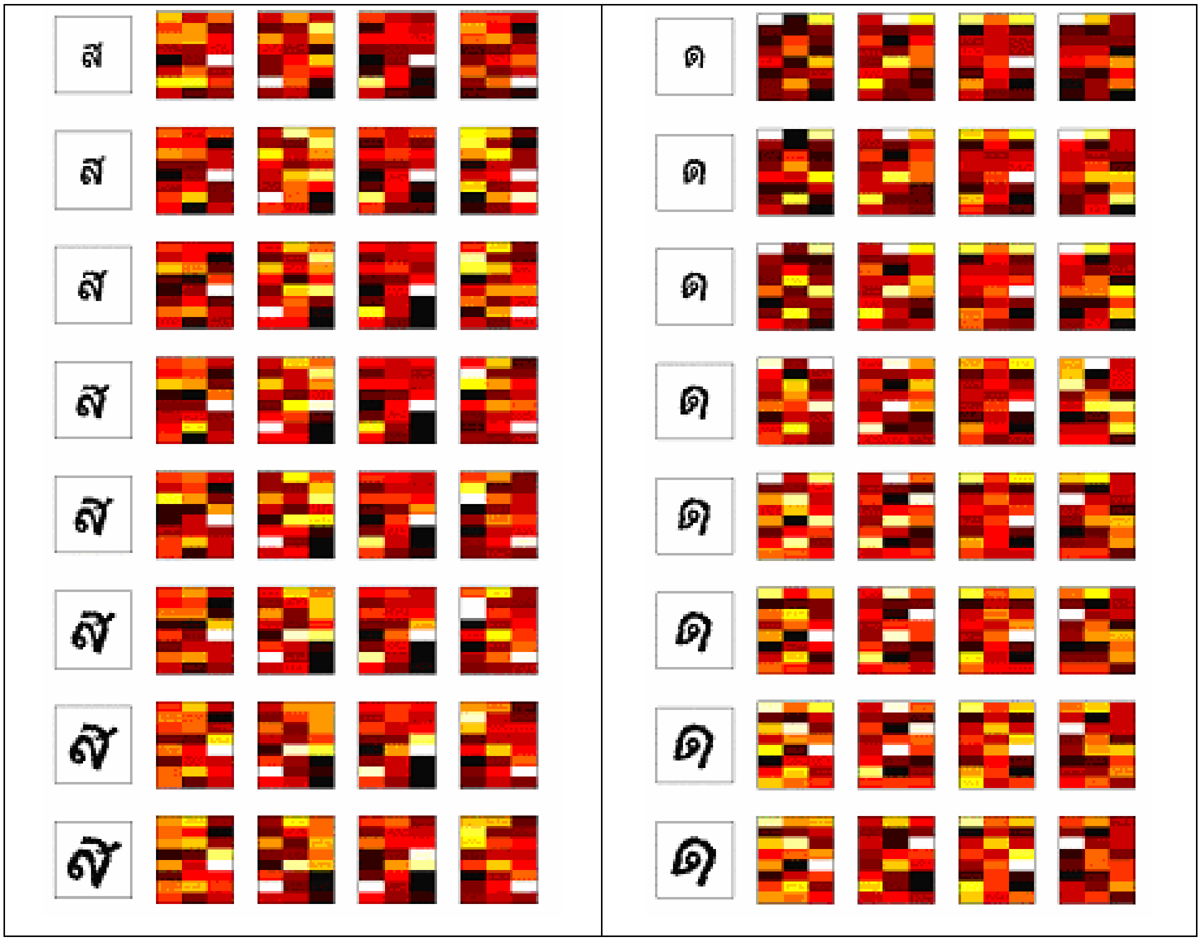
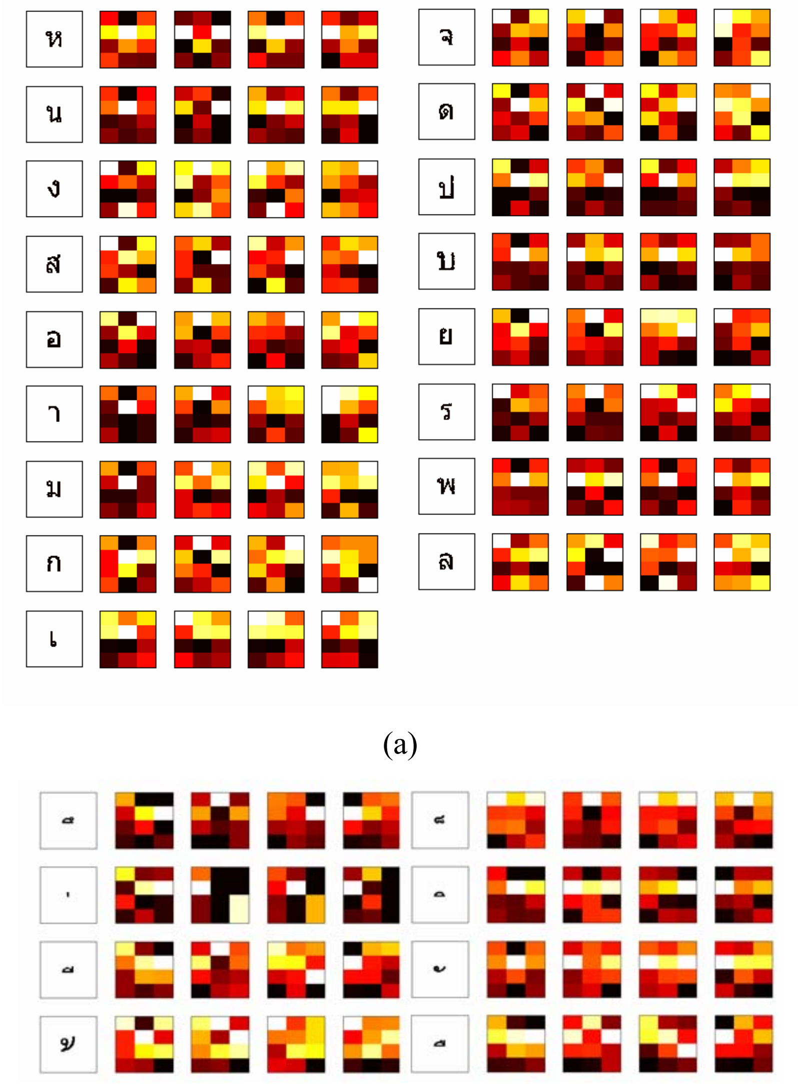


Figure 4.25 Variation of printed characters [ส] and [ด] with their corresponding edge direction features.



(a)

(b)

Figure 4.26 Edge direction features type I. (a) Characters written in body level. (b) Characters written at upper-level.

	ห	น	ึ	่	ง	ส	อ	า	ม	ี	้	ก	เ	จ	็	ด	ป	ิ	บ	ย	ร	พ	ั	ื	ล
ห	36	28	-	-	4	-	-	-	4	-	4	4	-	-	-	-	-	-	4	4	-	12	-	-	-
น	20	40	4	-	4	-	-	-	12	4	4	-	-	-	-	-	-	-	4	-	-	4	4	-	-
ึ	-	-	68	-	8	-	-	-	-	12	-	-	-	-	-	-	-	-	-	-	-	-	8	-	4
่	-	-	-	60	4	-	4	4	-	-	4	-	16	-	-	-	-	-	-	-	8	-	-	-	-
ง	4	4	-	-	44	4	4	4	-	4	-	-	8	4	-	-	12	-	-	-	4	-	-	-	4
ส	-	-	-	-	-	72	-	-	-	-	-	-	-	-	8	-	-	4	-	-	8	-	-	-	8
อ	4	-	-	-	4	-	56	4	-	-	-	-	4	8	-	-	-	-	-	8	4	-	-	-	8
า	4	-	-	4	-	-	-	68	-	-	-	-	8	4	-	4	-	-	-	-	8	-	-	-	-
ม	4	4	-	-	4	-	-	-	64	-	-	-	-	-	4	-	-	4	8	-	-	8	-	-	-
ี	-	-	4	-	-	-	-	-	-	68	-	-	-	-	-	-	-	12	4	4	-	-	-	8	-
้	8	-	-	-	8	-	-	4	12	4	44	-	-	-	-	4	-	4	-	-	-	-	12	-	-
ก	-	-	-	-	-	-	4	-	-	-	-	80	-	12	-	-	-	-	-	-	-	4	-	-	-
เ	4	-	-	20	4	4	-	4	-	-	-	-	60	-	-	-	-	-	-	4	-	-	-	-	-
จ	-	-	-	-	8	-	-	4	-	-	-	-	-	76	-	4	-	-	-	-	-	-	-	8	-
็	-	-	-	-	4	12	-	-	-	4	-	-	-	-	64	-	-	-	4	-	8	-	4	-	-
ด	-	-	-	-	-	4	-	-	4	4	-	16	-	-	-	68	-	4	-	-	-	-	-	-	-
ป	-	-	-	-	-	4	4	-	8	-	-	-	4	-	-	-	76	-	-	4	-	-	-	-	-
ิ	-	-	8	-	-	4	4	-	-	-	-	-	-	16	-	-	-	64	-	-	-	-	4	-	-
บ	-	8	-	-	-	-	-	4	4	4	-	-	-	-	-	-	-	-	68	8	-	-	-	4	-
ย	-	4	-	-	16	-	8	-	-	8	-	-	-	-	-	4	-	4	8	44	-	-	-	4	-
ร	-	-	-	-	-	8	-	4	-	4	-	-	-	4	4	-	4	-	-	-	72	-	-	-	-
พ	4	-	-	-	-	-	-	-	-	-	-	8	-	-	-	-	-	-	-	-	-	88	-	-	-
ั	-	4	-	-	4	-	-	-	-	-	-	-	-	-	-	-	4	-	-	4	-	-	84	-	-
ื	-	-	4	-	4	-	-	-	4	12	-	-	-	-	-	-	-	4	-	4	-	-	-	68	-
ล	-	-	-	-	-	16	-	4	-	-	-	4	-	-	-	-	-	4	-	-	4	-	-	4	64

Table 4.12 The confusion matrix for the classifier system using the edge direction type I. The average recognition rate is 63.84 percent.

	ห	น	ึ	่	ง	ส	อ	า	ม	ี	้	ก	เ	จ	็	ด	ป	ิ	บ	ย	ร	พ	ั	ื	ล
ห	60	8	-	-	-	-	-	-	-	-	4	8	-	-	-	4	-	-	4	4	-	8	-	-	-
น	12	32	-	-	-	-	8	-	-	-	4	4	-	-	-	-	4	-	4	-	-	28	4	-	-
ึ	-	-	48	-	-	-	-	-	12	16	-	-	-	4	-	-	-	8	-	-	-	8	-	4	-
่	-	-	-	76	-	-	-	4	-	-	4	-	16	-	-	-	-	-	-	-	-	-	-	-	-
ง	-	-	-	-	68	-	4	4	-	-	-	-	8	-	-	-	4	4	-	4	-	-	-	-	4
ส	-	-	4	-	8	40	8	-	-	-	4	-	-	4	4	-	-	-	-	-	4	-	-	-	24
อ	-	-	-	4	-	-	68	-	-	-	-	-	4	4	-	-	-	-	-	8	-	4	4	-	4
า	-	-	-	-	8	-	-	48	-	-	4	8	16	-	-	4	-	-	-	-	4	-	-	-	8
ม	4	4	-	-	-	4	-	-	64	-	-	-	-	-	-	-	4	-	8	-	-	12	-	-	-
ี	-	-	4	-	4	-	-	-	4	60	4	-	-	-	-	-	-	-	-	-	-	-	4	16	4
้	16	4	-	-	-	4	8	-	12	-	32	-	-	-	-	-	4	-	4	-	-	-	12	-	4
ก	-	-	-	-	-	-	-	4	-	-	-	60	-	4	-	16	-	4	-	4	-	4	-	-	4
เ	-	-	-	4	-	8	4	8	-	-	-	-	72	-	-	-	-	-	4	-	-	-	-	-	-
จ	4	-	-	-	12	-	-	8	-	-	-	-	-	44	-	-	-	-	4	4	-	4	-	-	20
็	-	-	-	-	4	24	8	-	-	-	-	-	-	-	40	-	-	-	8	4	-	4	-	4	4
ด	-	-	-	-	4	-	-	4	-	-	8	16	4	-	-	44	-	-	-	4	-	8	-	-	8
ป	-	4	-	-	-	-	-	-	4	-	-	-	4	-	-	-	72	-	4	8	-	-	-	4	-
ิ	4	-	4	-	-	4	-	-	-	-	4	-	4	-	-	-	-	64	-	-	-	-	-	-	16
บ	-	8	-	-	-	-	-	-	-	4	4	-	-	-	-	-	4	-	72	-	-	8	-	-	-
ย	-	-	-	-	-	4	4	-	12	-	-	-	-	-	-	-	4	-	16	48	4	-	-	-	8
ร	-	-	-	4	16	4	-	-	-	-	4	-	-	-	-	-	-	-	-	-	72	-	-	-	-
พ	-	4	-	-	-	-	-	-	8	4	-	4	-	-	-	-	-	-	-	-	-	80	-	-	-
ั	-	-	-	-	4	-	-	-	-	-	12	-	-	-	4	-	4	-	-	4	-	16	56	-	-
ื	4	-	-	-	4	-	-	-	8	16	-	-	-	-	-	-	-	-	4	-	-	-	-	64	-
ล	-	-	-	-	-	12	8	-	-	-	-	4	4	4	-	-	-	-	-	4	-	16	-	-	48

Table 4.13 The confusion matrix for the classifier system using edge direction type II. The average recognition rate is 57.28 percent.

4.9.4 The cavity image

Cavity is defined by a region bounded with at least three strokes (Gader el at 1984). The cavity types can be classified into 6 types: *North*, *East*, *South*, *West*, *Center* and *Hole* as illustrated in Figure 4.27. Cavity types are named by the open side of the cavity. The *Center* is bounded by four sides and the *Hole* is surrounded inside the drawing stroke.

The cavity properties can be used as the enhanced images with more useful information than a normal binary image or by transforming the cavity image into a feature vector. The number of cavity pixels in each normalized-size cavity image are counted for each zone then normalized to be ranked from 0 to 1. Table 4.14 shows the recognition result using the cavity image as the feature. The experiment setups are the same as previous feature evaluations.

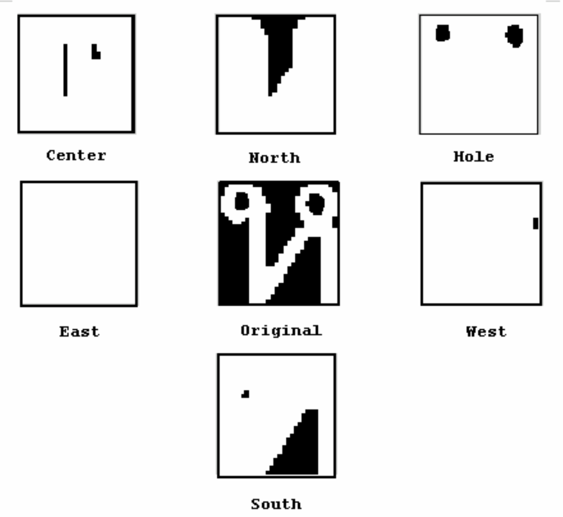


Figure 4.27 The cavity images of character [ฟ].

	ห	น	ึ	่	ง	ส	อ	า	ม	ี	้	ก	เ	จ	็	ด	ป	ิ	บ	ย	ร	พ	ั	ื	ล
ห	36	8	8	-	-	-	4	4	-	4	-	8	4	4	4	4	-	4	-	4	-	-	-	-	4
น	16	4	12	-	-	8	-	-	-	-	4	-	12	4	4	-	8	12	-	4	-	4	8	-	-
ึ	4	4	56	-	-	-	12	-	12	-	-	-	-	4	-	4	-	-	-	4	-	-	-	-	-
่	-	4	4	48	4	4	-	-	-	-	4	-	8	4	8	-	-	8	4	-	-	-	-	-	-
ง	4	-	4	-	40	-	12	-	-	-	4	-	4	4	8	-	-	8	4	-	4	-	-	-	4
ส	-	-	-	-	-	28	-	-	-	4	4	-	12	-	28	8	-	4	-	-	4	-	4	4	-
อ	-	-	-	4	4	-	64	4	-	-	-	-	-	-	4	8	-	-	4	8	-	-	-	-	-
า	4	-	-	8	-	-	-	68	-	-	-	8	-	-	-	-	-	8	-	-	4	-	-	-	-
ม	4	12	-	4	4	-	4	-	24	4	4	-	4	-	-	4	4	-	-	16	-	-	4	8	-
ี	4	4	4	-	4	4	-	-	-	36	12	-	4	-	-	4	-	4	-	-	-	4	-	12	4
้	4	-	4	4	-	-	-	-	-	-	32	4	8	-	-2	4	4	-	-	-	4	8	4	-	-
ก	4	-	-	-	-	-	4	4	-	-	4	60	-	-	4	16	-	-	-	4	-	-	-	-	-
เ	4	-	4	12	-	-	-	-	-	-	4	-	56	4	4	4	-	-	-	-	4	-	4	-	-
จ	-	-	-	-	-	-	-	12	-	4	-	4	-	56	-	4	-	8	-	-	8	-	-	4	-
็	4	-	-	8	-	12	8	-	-	-	4	4	-	-	44	4	8	-	-	-	-	-	4	-	-
ด	-	4	-	8	-	-	4	4	4	-	-	16	4	-	4	52	-	-	-	-	-	-	-	-	-
ป	-	-	4	-	-	4	-	-	-	-	8	-	12	-	8	8	28	4	4	4	-	4	4	-	8
ิ	-	-	-	4	-	4	-	12	-	8	-	-	8	8	-	4	-	28	-	-	4	8	-	4	8
บ	16	-	4	4	4	-	4	-	8	-	-	4	-	-	4	4	8	-	20	8	-	8	-	4	-
ย	-	-	8	-	-	-	8	-	-	-	8	-	4	-	-	4	-	4	4	52	-	8	-	-	-
ร	-	-	-	-	4	4	-	8	-	-	4	4	12	-	4	4	-	-	-	-	56	-	-	-	-
พ	8	-	4	4	-	-	-	-	4	8	-	-	-	-	-	-	8	-	-	-	-	-6	4	-	-
ั	-	-	8	4	-	-	4	-	-	-	4	-	4	-	-	-	8	4	-	4	-	4	56	-	-
ื	4	-	4	-	-	-	-	-	-	4	-	-	-	-	-	-	-	4	4	-	-	-	-	80	-
ล	-	-	-	-	-	8	8	-	4	-	4	-	4	-	-	4	4	4	-	4	-	8	-	-	48

Table 4.14 The confusion matrix for the classifier system using cavity image feature. The average recognition rate is 45.28 percent.

4.10 Summary

This chapter introduces the data collection process for both handwriting characters and words for the legal amount used for training and testing purpose. Basic statistical properties of the data are analyzed and some problematic characteristics of Thai handwriting character are illustrated to show the difficulties of the problem. A set of selected features are applied to the printed character image to demonstrate the output characteristics. These features are then used to train the neural networks to classify the handwriting characters. A small set of handwriting data has been prepared for this evaluation purpose. The results of the recognition indicate that each feature does not work well to recognize the handwriting data. In the next Chapter, the recognition system to improve the recognition rate for Thai handwriting in legal amount is proposed.

Chapter 5

Thai Legal Amount Handwriting Recognition

In the previous Chapters, the preliminary experiments indicate that simple recognition systems have a low recognition rate which is impractical in legal amount recognition applications. This study proposes three techniques to solve the problem. The first technique is to improve the character recognition rate at the preprocessing and the recognition process stages, which depend on information from the characters' shapes. The second technique is to include post processing processes, which apply a prior knowledge of the lexicon and certain rules of syntax for the legal amount to improve the overall recognition rate. The last technique is to reduce the effect of irregular characters by locating the possible word boundaries from the irregular characters positions. The hypothesis word images are segmented and recognized by word recognition with the holistic approach. These words are then used to form amounts and checked with the syntax rules again. In the final stage, the highest score choice of all syntax checked amounts will be selected as the answer.

5.1 Improved preprocessing and recognition process

High character recognition rate is the foundation for overall system performance. Simple character recognition systems with single features have the highest recognition rate at 60 percent. This rate is not adequate for the legal amount recognition applications such as check reader systems, since it requires very high recognition rate, i.e., 95 to 100 percent with some rejection rates. To improve the character recognition rate, this Chapter proposes techniques; that include dividing characters into subclasses, using multi-feature combinations and using classifiers that indicate the rank of possible characters. Figure 5.1 shows the diagram of the improved character recognition systems.

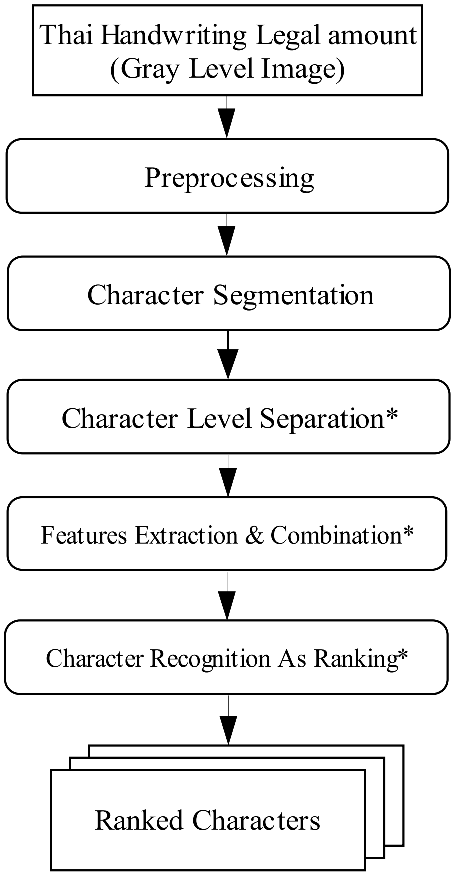

Thai Handwriting Legal amount
(Gray Level Image)
Preprocessing
Character Segmentation
Character Level Separation*
Features Extraction & Combination*
Character Recognition As Ranking*
Ranked Characters

Figure 5.1 The diagram of the improved character recognition process. The processes with * are the techniques to improve the recognition rate.

The amount images are scanned with a flat-panel scanner into grey level images. The preprocessing processes are image-processing processes to improve the image quality and convert the image into a suitable format before passing it to the feature extraction process. The images are originally scanned at 256 grey-levels. The grey-level images are converted into binary images using Otsu algorithm (Otsu 1993). However, Thai character segmentation for touching or overlapping characters is still under investigation. There is no solution for this problem yet. In this study, the connected characters are treated as single characters.

5.1.1 Character writing level group separation

In a recognition system, a large class can be divided into subclasses with smaller members. Different recognition systems with different features can be selected to be the best classifier for each subclass. These concepts can raise the recognition rate, given that the inputs are separated into correct subclasses. For Thai legal amount, the characters can be divided into two subgroups by their writing level before the recognition process to improve the recognition performance.

The standard Thai writing system has four writing levels, but the legal amount words are drawn using only three writing levels, namely the body, vowel and upper-vowel level. The consonants are always drawn at body level and most vowels are drawn at vowel or upper-vowel level. Three vowels drawn at the body level are [เ], [แ] and [า]. There are eight vowels drawn at the vowel and upper-vowel level. These characters are grouped as the high group characters. There are seventeen characters drawn at body level. These characters are grouped as the body group characters. Table 5.1 shows two groups of characters based on their writing level.

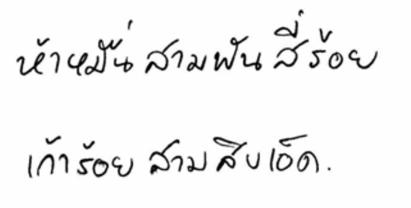
High group	ึ	่	ื	็	ี	้	ั	ิ	
Body group	ห	น	ง	ส	อ	า	ม	ก	เ
	จ	ด	ป	บ	ร	ย	พ	ล	

Table 5.1 Characters grouped by writing level.

The success of the idea to increase recognition rate depends on algorithms to separate the input into subclasses correctly. If these characters are separated into incorrect subclasses, the recognition results of those characters are definitely wrong. The character writing level group separation algorithms are implemented based on the fact that all vowels drawn at the vowel and upper-vowel level are shorter than the consonants and these vowels must be drawn above the consonants. Figure 5.2 shows an example of legal amounts in printed and handwriting to illustrate this fact.

ห้าหมื่นสามพันสี่ร้อย
เก้าร้อยสามสิบเอ็ด

(a)



(b)

Figure 5.2 Legal amount of ห้า-หมื่น-สาม-พัน-สี่-ร้อย (53,400) and เก้า-ร้อย-สาม-สิบ-เอ็ด (931) (a) Printed form (b) Handwritten form.

The printed characters, both consonants and vowels, are placed in a line with almost fixed and specific positions. Therefore, it is very simple to implement algorithms to separate the printed character based on their positions. However, handwriting characters are drawn within some more flexible boundaries rather than fixed positions. The lower bound of some vowels are always found very close or overlapped with the upper bound of the consonants and characters can be seen as they are drawn based on some imaginary line which moves up or down along the writing line. However, the algorithm to separate the handwriting characters into two writing levels can be based on their height and their positions described in the following steps.

First, the noise that appears as small dots or small objects must be removed. This can be done by finding the average area of every component in the binary image. A threshold is selected with respect to the average area, i.e., ten percent of the average area. Any object smaller than the threshold will be removed. The second step is to find the components that seem to be shorter than most of the characters. Then find any components higher than a half of the average height of all objects and group them in *body* character group. The *high* character group is supposed to be drawn above the *body* character group so any components of which their lowest positions are above average height will be grouped in the *high* group. Finally, assign the remaining objects to the *body* group. The block diagram of the algorithm is shown in Figure 5.3.

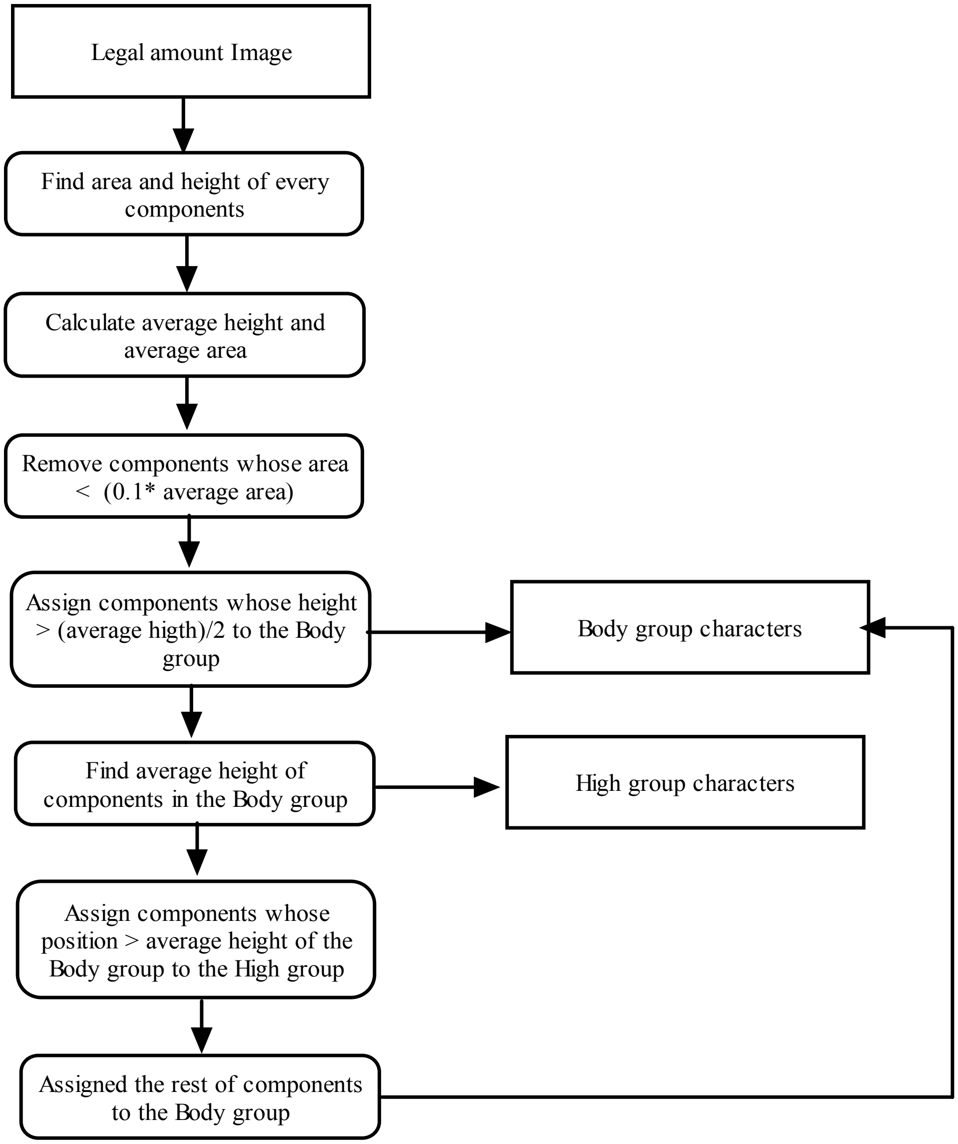

Legal amount Image
Find area and height of every components
Calculate average height and average area
Remove components whose area < (0.1* average area)
Assign components whose height > (average higth)/2 to the Body group
Body group characters
Find average height of components in the Body group
High group characters
Assign components whose position > average height of the Body group to the High group
Assigned the rest of components to the Body group

Figure 5.3 The character writing level separation algorithm.

5.1.2 Multiple features combination

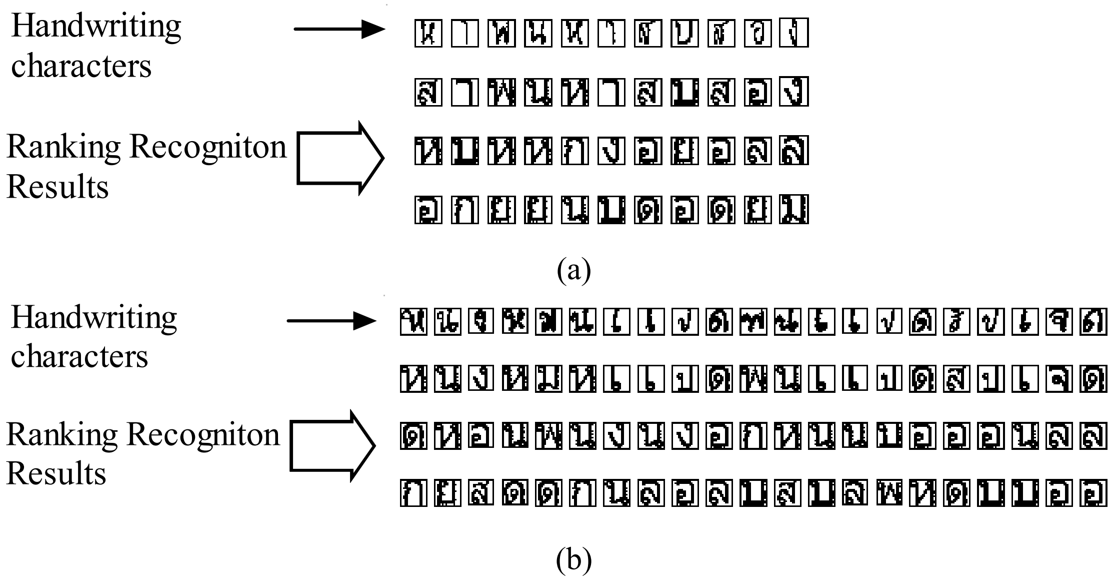
In the previous Chapter, features that could be used for the characters recognition are Fourier descriptors, projection profiles, edge direction, and topology images. These are evaluated for if any of them could perform well for the handwriting character recognition. However, the results indicate that when using each feature with a simple classifier, the recognition rate is considerably too low for the legal amount recognition applications. The second technique to improve recognition rate is using multiple features. Each feature could be good for some specific cases or forms of characters. These features can be combined together to capture more forms and classes of the handwriting characters and the recognition rate should be raised. However, the combination of multiple features might be very redundant and require a larger size of samples. Furthermore, increasing the number of feature vectors could lead to an increased recognition rate but this is not guaranteed. The empirical experiments are needed to find a right combination of the features.

The size of combining features vector can be reduced by Principle of Component Analysis (PCA) transformation to represent large dimension feature vectors with smaller dimension vectors. The PCA transforms the input vector uncorrelated with each other, orders the feature vector to the largest variation and eliminates the least variance contribution components (Hagan et al 1998). The details of how to use PCA to reduce feature vector's dimension are discussed in Duda et al (2001).

5.1.3 Character recognition systems with the ranked results

The recognition systems are created by the training process coupling with the training data set and the preferred targets or classes. The training objective is to adapt the system with given input to generate output similar to the preferred target. After the training process, the recognition systems are expected to generate answers to the input that are not in the training data set. Conventional recognition systems select the highest score output as the answer for a given input data or a feature vector.

For the handwriting character recognition problems, a single answer might not be the most desirable because a similar shape can be shared between different characters, especially in the multi-writers environment. It is preferable to have recognition systems that give results as the possible answers with confident score or ranked results. Instead of a single answer, the results are the ranked of top *n* highest score. The probability that the answers could cover the correct character is increased, as is the recognition rate. In this study, the top three classes of the highest scores are selected. Multilayer Layer Perceptron and *k*-NN classifiers are often applied to character recognition problems. Both can generate ranking answers. However, the ranking answers cannot be used directly. The post processes are needed to select and combine these answers into the useful forms. In handwriting legal amount recognition application, the ranking characters are joined to form the amount words and matched with the lexicon to find all possible words. Figure 5.4 shows some examples of the results of ranking characters. The first line is the actual handwriting characters and the following lines are the recognition results.



(a)

(b)

Figure 5.4 Examples of character recognition results. (a) The first amount is ห้า-พัน-ห้า-สิบ-สอง (b) The second amount is หนึ่ง-หมื่น-แปด-พัน-สิบ-เจ็ด.

5.1.3.1 Multiple Layer Perceptron with ranked output

The Multilayer Layer Perceptron (MLP) is used successfully in classification problems. General MLP classifier systems are made of three layers: the input layer, the hidden layer and the output layer. Size of the input layer is equal to the size of the feature vector. The size of the output layer is equal to the number of output classes. The size of hidden nodes is normally between the number of input and output nodes.

The MLP are trained with the training set with the desired output. The output coding is produced by setting one value to the desired output and zero value for the other outputs. Designing an MLP with ranked answers is possible through two approaches. The first approach uses the conventional training with one value to the desired output but when using the classifier all *n*-th highest score classes are selected. The second approach use the confident class value of every data set and then trains the MLP with these values. Then all *n*-th highest score classes are selected as the previous approach. Given an image of a character S, and set of character classes C_i, the character confidence assignment is a process that assigns a value to a character S indicating the degree to which S represents C_i. The degree ranges from 0 to 1 corresponding to the least and the most confident. A character that assembles to two classes can have equal values of confident. There are several ways to assign confidence to each character. The simple way is by assigning the confidence value calculated from the clustering analysis of the features of characters.

5.1.3.2 *k*-Nearest Neighborhood Classifier

The *k*-Nearest Neighborhood Classifiers (*k*-NN) are among the simplest classifiers that work well with a large number of training samples. The answer of *k*-NN classifiers depends on the majority of *k* nearest neighbor samples around the unknown input. As the number of *k* gets larger, the error rate of this classifier is closed to the Bayes error rate. In practice, *k* is empirically selected to be large enough to obtain a reliable rate. The data set for the classifiers must be tagged with their class and the classifiers keep all tagged data in the system without the training process. Both their labels and distances are useful for calculating the confidence of the choices. Figure 5.5 illustrates the *k*-NN classifier mechanism. The two groups of data are shown as diamond and triangle shape points. The unknown class is at point X. A spherical region is grown from that point until it contains *k* points. The class assigned to input X is the result of a majority vote of its neighbor points.

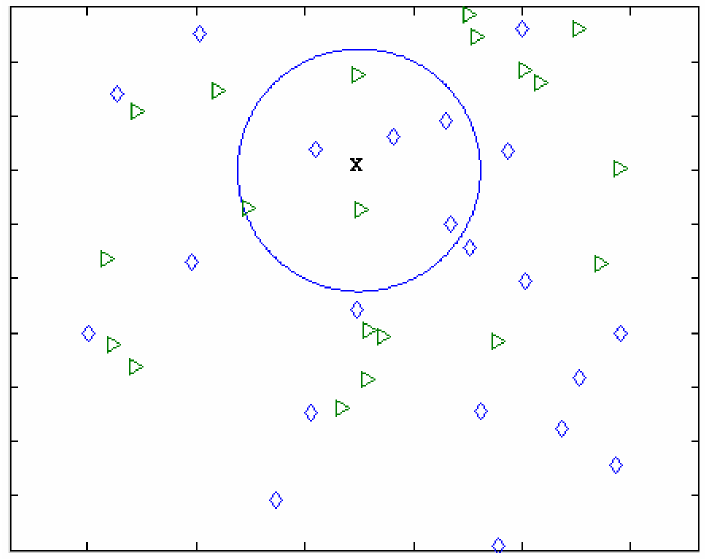


Figure 5.5 The *k*-NN classifier, k equals to 7.

The k-NN classifiers consider the neighbor by the distant function. In this study, the Euclidean distance is used as the metric function, assuming all features are normalized. The result of *k*-NN classifiers gives a rank of candidates with scores instead of one specific class above threshold, as general MLP does. Figure 5.6 shows examples of the classifier results with characters [ห] and [น].

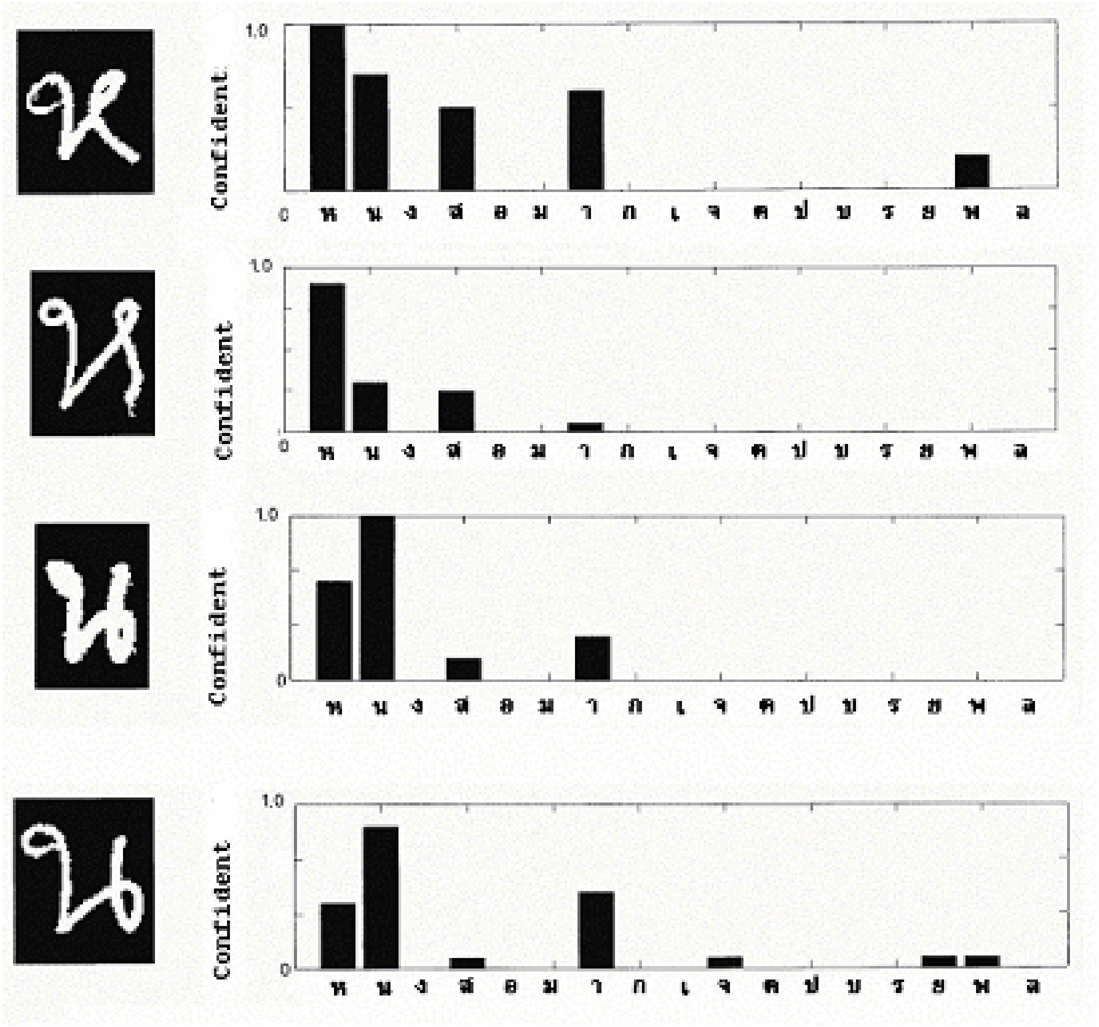


Figure 5.6 Character [ห] and [น] and the ranked results of *k*-NN classifier.

5.2 Post processing

If handwriting character recognition processes could identify every character perfectly, no further processing would be necessary. However, many cases have proved that information provided at character level alone is not enough to identify the character class. The second necessary component to improve the recognition rate is the post processing processes, by applying a prior knowledge of the lexicon and certain rules of syntax for the legal amount. First, the results of character recognition in ranked answers are searched for whether any sequence matches any possible words of the legal amount lexicon. These possible words are then joined together to form choices of amount called *amount paths*. Only words that continue can be joined together. The complete *amount paths* are amounts that fit along the full length of the input sequence without interruptions. Those that fail this condition are called *incomplete paths*. If any complete amount path can pass the syntax check, which means some legal amount are found, there is no used for the incomplete amount paths. These complete amounts are checked as to whether they comply with syntax rules. Any amount paths that conflicted with the syntax rules are rejected. Figure 5.7 shows the diagram for the post processing. However, if no complete amount paths can pass the syntax check, the input characters should be checked to find any anomaly size or position that might create some recognition error. This idea will be used in the third part of the proposed techniques.

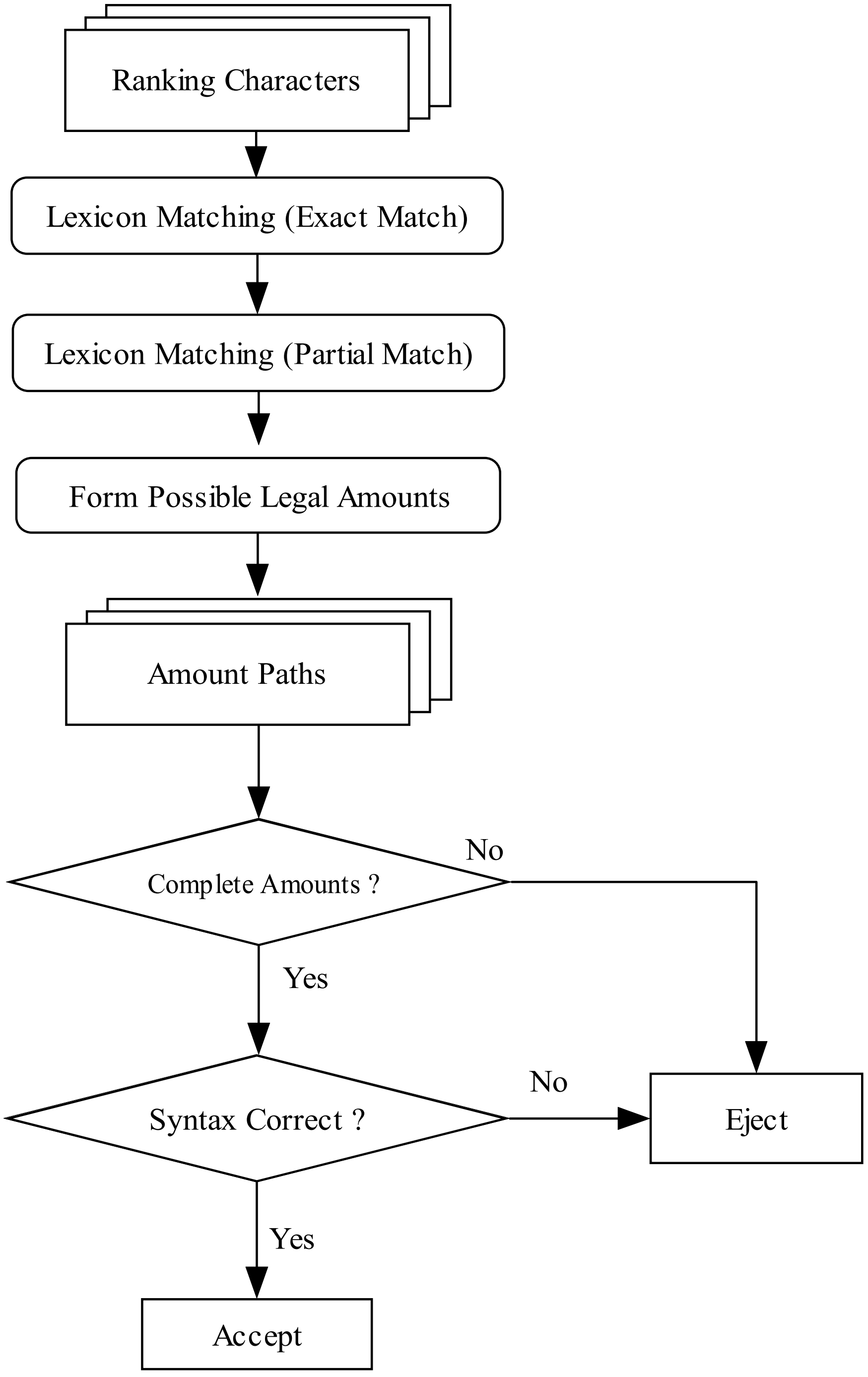

Ranking Characters
Lexicon Matching (Exact Match)
Lexicon Matching (Partial Match)
Form Possible Legal Amounts
Amount Paths
Complete Amounts ?
No
Yes
Syntax Correct ?
No
Eject
Yes
Accept

Figure 5.7 Post processing processes for the legal amounts recognition.

5.2.1 Lexicon matching

If character recognition systems work perfectly, all recognized characters are supposed to be matched with the words in lexicon, given that characters are in sequence. However, Thai handwriting contains multiple levels and the location of vowels above some consonant might be leading or lagging from the proper positions. When characters are segmented from the image, the sequence of vowels and characters are shuffled, i.e., the vowel above consonant may come before or after the consonant. When a word is written with three levels, the vowels are sometimes away from the proper positions from one to two positions. The word matching algorithms must to be very flexible to handle the possibly shuffled input characters sequence. Another solution is derived from the fact that legal amount words can be fully identified with only information from the *body* group characters, for example, a sequence of [หนง] would recall to the word หนึ่ง and a sequence of [หมน] would recall to the word หมื่น. Table 5.2 shows all amount words and body group characters of those words.

Original	หนึ่ง	สอง	สาม	สี่	ห้า	หก	เจ็ด	แปด	เก้า
Body only	หนง	สอง	สาม	ส	หา	หก	เจด	แปด	เกา
Original	สิบ	ร้อย	พัน	หมื่น	แสน	ล้าน		ยี่	เอ็ด
Body only	สบ	รอย	พน	หมน	แสน	ลาน		ย	เอด

Table 5.2 The original amount words and body group character of those words.

The matching algorithm using only *body* group characters is much simpler than the matching of a possible shuffled sequence. This solution is used in this study. In addition, the vowels in upper level can be used for confirmation. When the words are joined together to form a complete legal amount, information provided from the *body* group characters is still enough to reveal the original words correctly. Table 5.3 shows some examples of legal amounts where high group characters are removed and their corresponding original legal amounts.

***Body* characters only**	**Original legal amount**
หนงรอยเกาสบเอด	หนึ่งร้อยเก้าสิบเอ็ด
สามหมนสพนหารอยเกาสบสอง	สามหมื่นสี่พันห้าร้อยเก้าสิบสอง
สามแสนสหมนสพนสามสบเอด	สามแสนสี่หมื่นสี่พันสามสิบเอ็ด
สามพนหกรอยสสบส	สามพันหกร้อยสี่สิบสี่

Table 5.3 The examples of sequences characters of *body* group characters and the original legal amounts.

However, the ranked character results imply that these answers contain some errors. In case of the accumulative top-3 ranked answers, the best scenario is that one of the three answers is correct and the other two are wrong. In the worse case, all three characters are wrong. The matching processes only reveal some words that might be possible but not always correct. To verify if they are correct, these words are joined together to form choices of legal amounts called *amount paths*. The *amount paths* will be checked if they obey the syntax rules in the final step. There are two matching methods: the exact match and partial match. The exact match uses the full word to search possible matches with input strings while the partial match uses part of the word.

5.2.1.1 Exact match process

The exact match algorithms locate any match of words in lexicon in the input sequence. The algorithms first select the longest word, i.e., แปด or แสน, to match with input sequence and followed by the shorter words. Figure 5.8 demonstrates the simplified exact match operation to a character sequence. The input characters sequence is [สองพนหารอยสบ], given a limited set of lexicon of [สอง], [รอย] [พัน] [หา] and [สบ] as shown on the left. Assuming that there are no errors on the character recognition, the exact match can reveal all words in the amount. For Thai legal amount lexicon, character [ส] and [ย] are excluded in the exact match process because they are only one *body* character and are likely to be confused with other words. These characters are included in the partial match process instead.

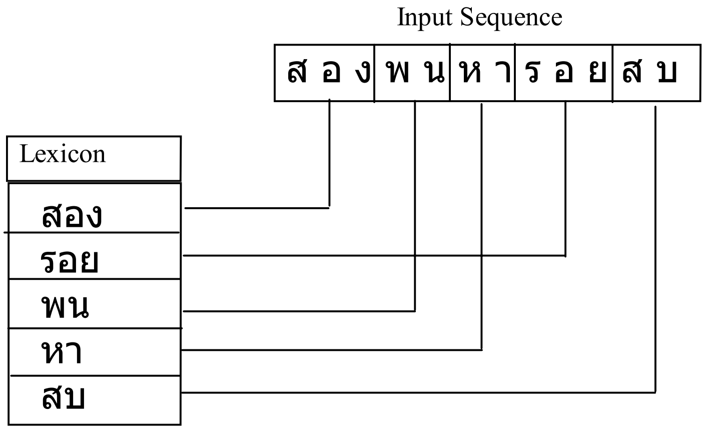


Figure 5.8 The exact match operations for a character sequence.

However, the output from character recognition is a ranking result. One handwritten character can have *n* possible answers. The matching algorithm must search though *n* possible choices per character. Figure 5.9 shows the exact match operations with the three ranked character recognition results of the handwriting of ห้า-พัน-ห้า-สิบ-สอง (5,052). The handwriting is shown in the first line, the ranked recognition results are shown below, and the matching results are shown in the third part. There are nine amounts matched at six different positions.

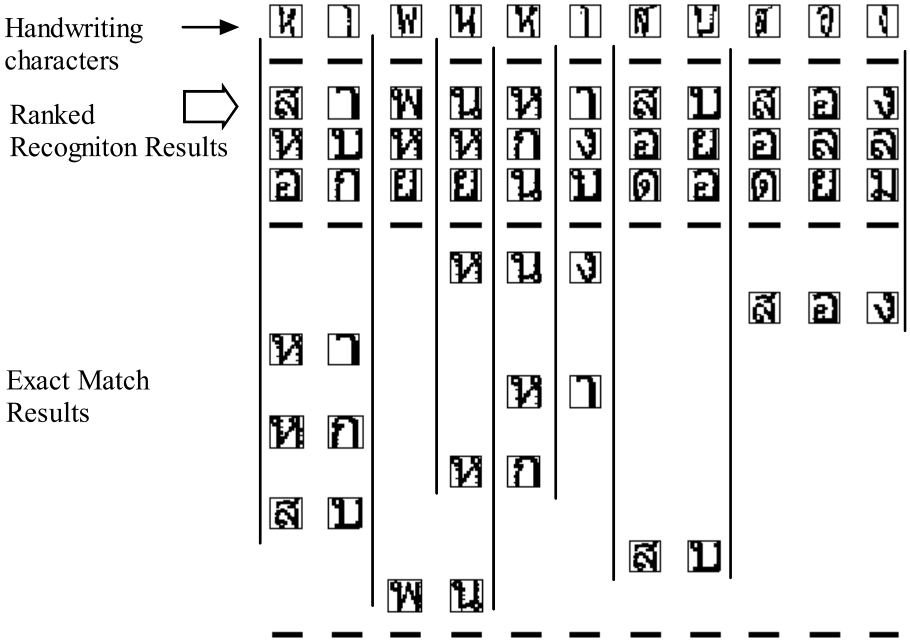


Figure 5.9 The exact match for ranking input of ห้า-พัน-ห้า-สิบ-สอง.

5.2.1.2 Partial match process

The partial match assumes that there are some errors in character recognition. In this study, one replacement error per word is used for the match, thereby, allowing one misrecognition character per word. The longest word is used first to search any partial match with the sequence followed by shorter words. A word of n character's length has n possible patterns of string matched. The partial match result might be redundant and overlap with others. The partial match compares the input string with words in the lexicon with more robust results. The demonstration of partial match is shown in

Figure 5.10. There are four errors introduced in the input sequence. Each error only exists in one word. The partial match can reveal the possible words in the input sequence.

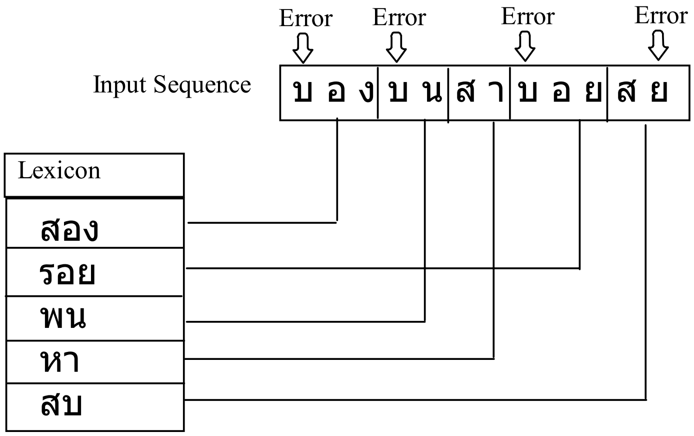


Figure 5.10 The partial match operation.

The partial match for the ranking results works in the same way of the exact match. All possible characters in each position must be searched and compared with words in the lexicon, given that one replacement error is allowed. The number of possible matches might be very large for a long amount word. Figure 5.11 shows the result of a partial match for amount ห้า-พัน-สิบ-สอง.

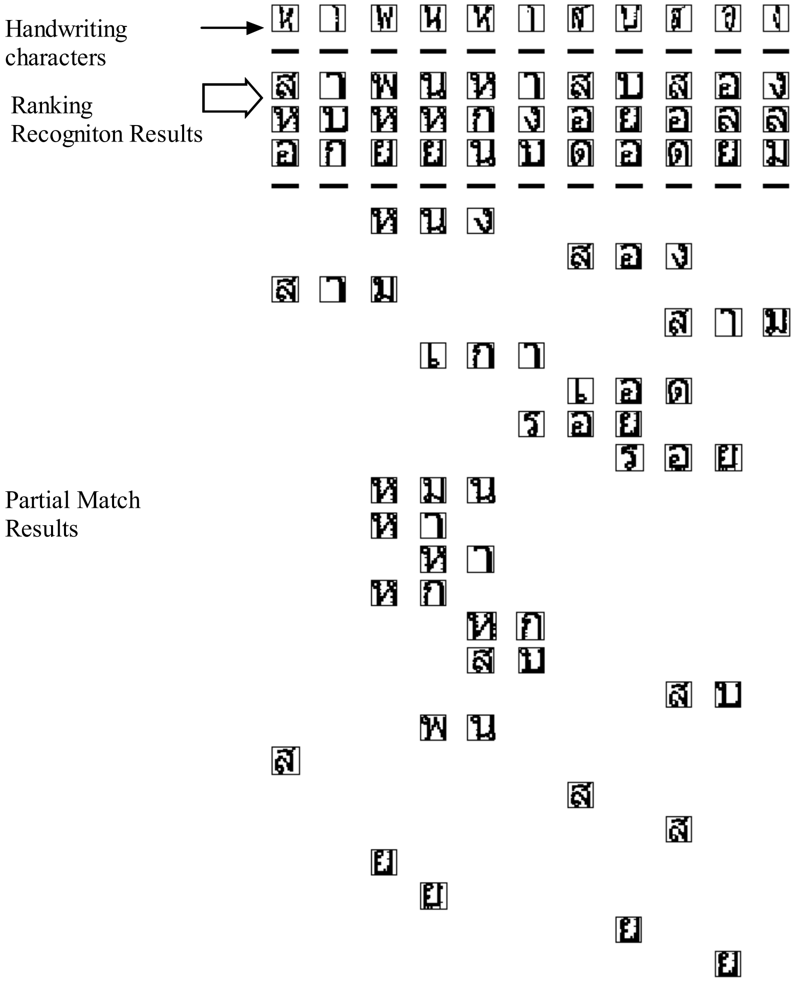


Figure 5.11 The partial match results.

5.2.2 Forming possible choices of legal amount

Both exact and partial match results must be integrated together to form choices of possible amounts, named *amount paths*. Each matched word has its starting and ending position. To join two words together the second word must have the starting position next to the ending position of the first word. For example, if the first word has the ending position at 5, only the words that have the beginning position at 6 can be joined. This is appears to create possible paths from one word to another until the last position of the input string. The complete *amount paths* are defined as the joining words that start for the first position of the input sequence and end at the last position of the input sequence without interruption or being broken. The incomplete *amount paths* are those which fail this condition.

Figure 5.12 shows the integration of exact and partial match results. At the first position of input sequence, the possible words are ห้า, หก, สิบ, สาม and สี่. If the word ห้า is chosen in the first step, the next possible word could be พัน, หนึ่ง, หมื่น, ห้า, หก and ยี่ because these word have the starting position next to the ending of word ห้า. If word พัน is chosen in the second step, the next possible words are ห้า, หก, and สิบ. If word ห้า is chosen in the third step, the next possible words are สิบ, สอง, เอ็ด and สี่. Some examples of complete *amount paths* are [หา พน หา สบ สอง], [หก พน หา สบ สอง], [สบ พน หา สบ สอง] and [หา หา หา สบ สอง]. They are the results of joining words, both exact and partial matching at A, B, C, D, E, F and G as indicated in the Figure. If the words in amount are all recognized or contain only one error per word, the complete *amount paths* would contain the correct answer. The correct answer can be revealed after the *amount paths* are checked with the syntax rules at the final step.

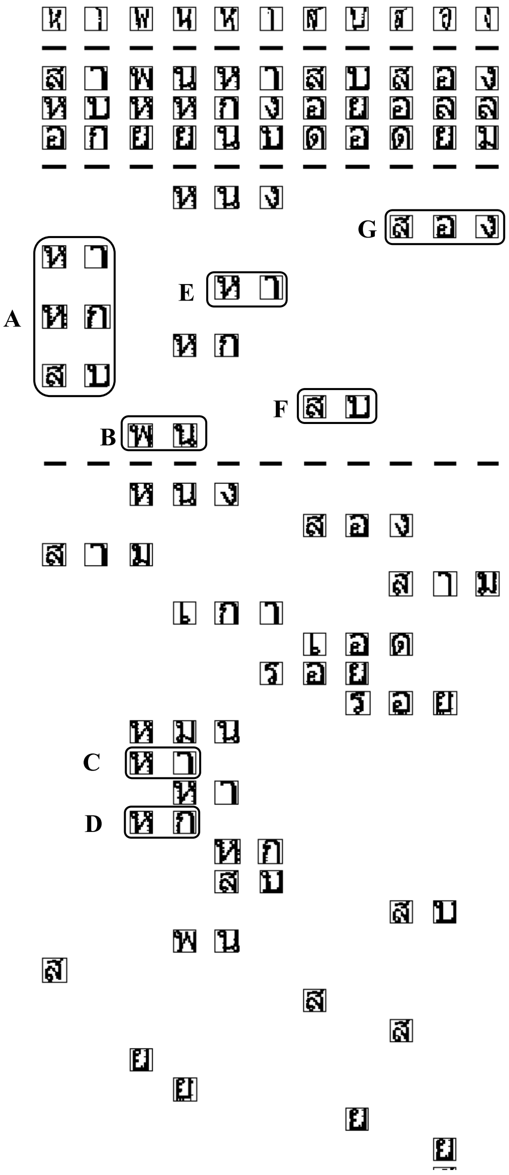


Figure 5.12 The integration of exact and partial match for ห้า-พัน-ห้า-สิบ-สอง.

5.2.3 Legal amount syntax

Syntax is a set of rules that govern the composition of words to express some meaning. It can be very complicated and consists of hundreds of rules. However, the syntax for the legal amount is only a subset of those for general language. Thai legal amount syntax is simple and requires a small set of rules. In this section, the syntax of Thai legal amount will be described and followed by the implementation of the algorithm. The syntax described in this section is for the legal amount of seven digit numbers only. For higher digits, some rules need to be changed. It should be noted here that syntax for Thai language is very flexible and very simple compared to those of English. The syntax rules described here are restricted to formal writing which require that the first word in legal amount must be a *numeric* word, except สิบ. The following examples are regarded as informal legal amount because they do not begin with a numeric word: ร้อยสิบ, พันห้าร้อย, หมื่นสาม and แสนสอง.

State diagrams for a legal amount up to 4 digits are shown in Figure 5.13. The A, B, C and are sets of words defined as the following:

A = {เอ็ด, สอง, สาม, สี่, ห้า, หก, เจ็ด, แปด, เก้า}
B = {ยี่, สาม, สี่, ห้า, หก, เจ็ด, แปด, เก้า}
C = {หนึ่ง, สอง, สาม, สี่, ห้า, หก, เจ็ด, แปด, เก้า}

2Ds, 3Ds, and 4Ds are the starting state for two, three and four digit amount, the state name is of the format nDs, where n is the digit and s is the stage. E is the terminated state. The transition only occurs when supplied with input as indicated over an arrow and the e symbol represents the last input.

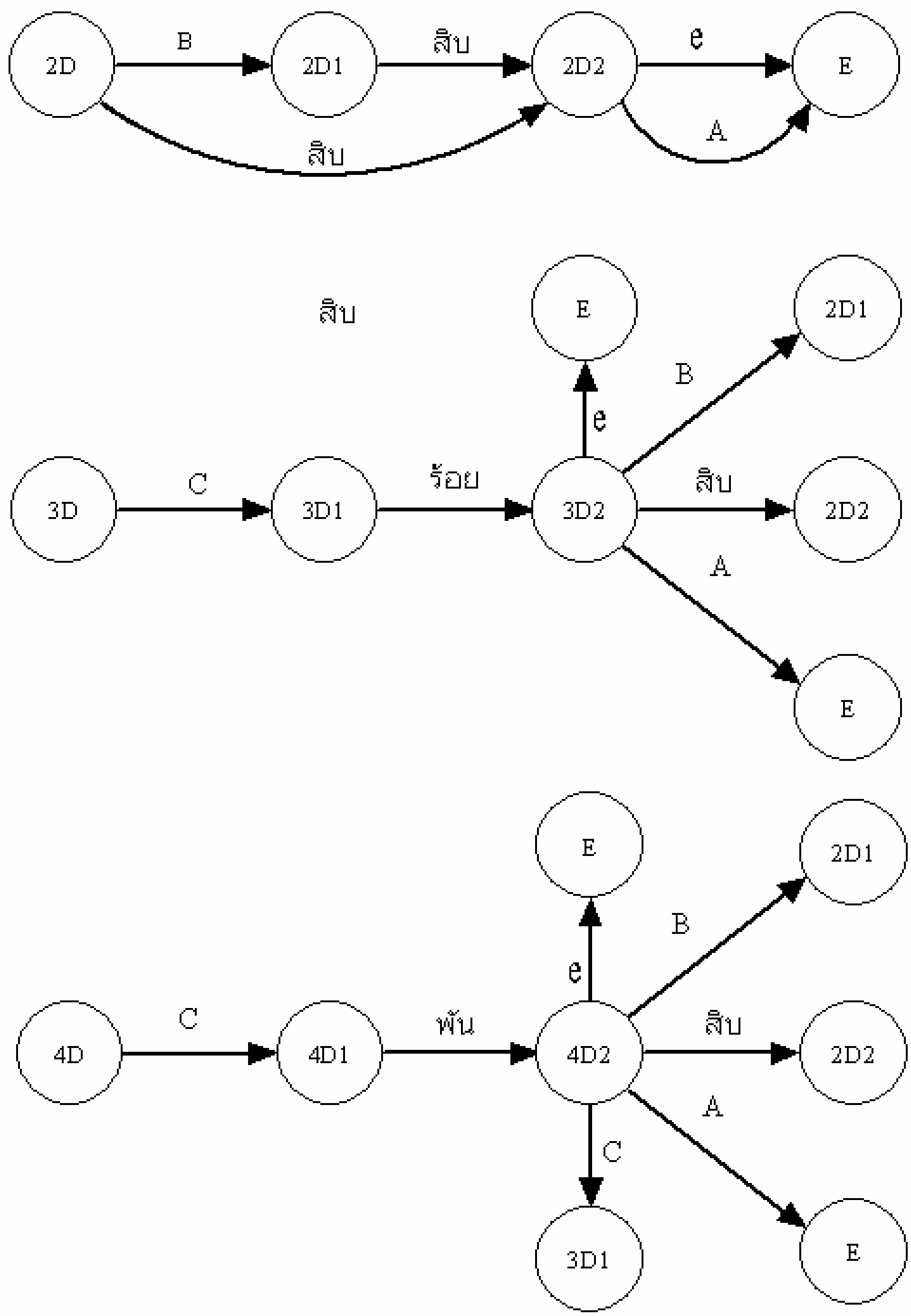


Figure 5.13 State diagram of two, three and four digit amounts.

Syntax rules for legal amounts can be divided into three groups. The first group concerns order of the *tenth-power* word. The order of the *tenth-power* words are sorted as ล้าน (1,000,000)-แสน (100,000)-หมื่น (10,000)-พัน (1,000)-ร้อย (100)-สิบ (10). Any mixed order words are not allowed, for example, สิบสี่ร้อย (สิบ-ร้อย sequence) is as the example of wrong syntax. Each *tenth-power* word can only exist once in an amount. For example, ห้าร้อยหกร้อย violates the syntax rules because the word ร้อย is written twice.

The second group concerns types of adjacent words. The adjacent words cannot be of the same type. For example, หกเก้า (*number-number*) and ร้อยพัน (*tenth-power-tenth-power*) are the wrong syntax. One exception is for word สิบ which can be put after other *tenth-power* words, i.e., หนึ่งร้อยสิบ (110), หนึ่งร้อยสิบเจ็ด (117), สามพันสิบหก (3,016), or สี่หมื่นสิบสอง (40,012). The word สิบ (ten-10) can lead by any *number* except หนึ่ง or สอง such as สามสิบ (30), สี่สิบ (40), ห้าสิบ (50), or หกสิบ (60) but not หนึ่งสิบ nor สองสิบ.

The third group concerns the position of words in legal amounts. The legal amounts always begin with *number* for example, หนึ่งล้านสามแสน ห้าพันสี่สิบเอ็ด unless the amount is a two digit from 10 to 19. These amounts can be written as สิบ (10) สิบเอ็ด (11) to สิบเก้า (19). The word เอ็ด is used only with amounts that have more than one digit and the least significant digit of that amount is 1, such as สามสิบเอ็ด (31). The word เอ็ด exists at only the last position of amount. The word ยี่ is used only with amounts more than one digit and 2 is the second digit of the amount. It is always followed by สิบ to make the complete phrase as ยี่สิบ (20).

5.3 Recovering processes

The previous section explained procedures that should be able to handle the handwriting when each character is drawn separately, and characters recognition results have no more than one error per word. However, handwriting will always contain connected characters that the simple characters segmentation fails to separate into single characters. This section describes procedures to recover the words having connected characters.

Irregular characters in handwriting such as connected or broken characters might cause input character sequences shorter or longer than the actual length. Therefore, recognition answers to these characters are always wrong. After the lexicon matching, the joining of these matched words often leads to incomplete amounts, or complete amounts that are rejected by the syntax check after all. If the shape of the word that contains the connected characters retains most of the word feature, then this word image could be recognized by the word recognition.

Figure 5.14 shows the recovering processes to recognize the incomplete amounts. First, some irregularities in handwriting are located and patterns of word segmentation are performed to segment possible word images from the image of the full amount. These images are just the hypothesis words that needed to be verified. The recognition of these hypothesis words are joined with the previous character recognition results to form the amounts paths again and have the syntax checked as in the previous procedures.

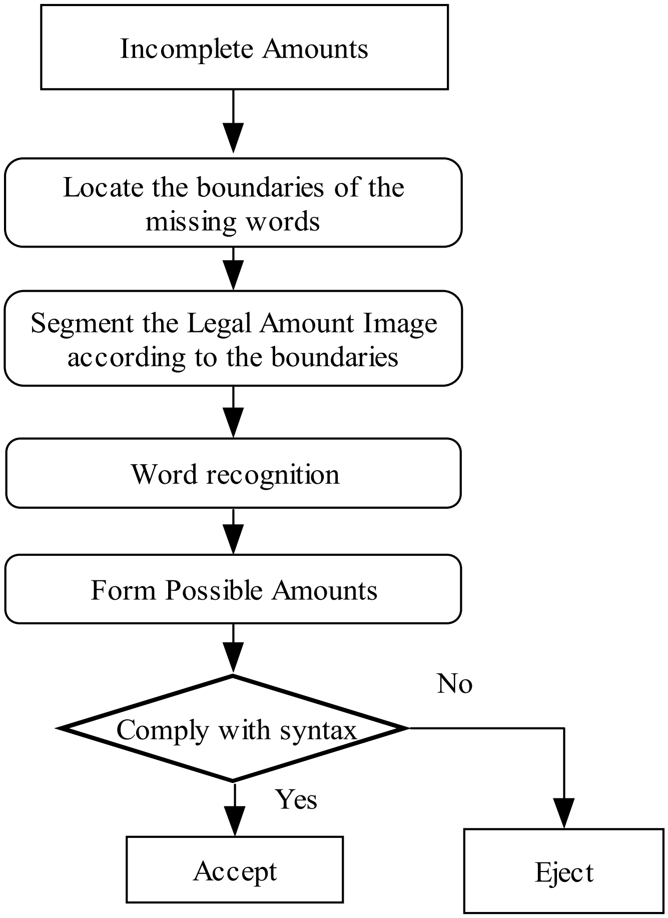

Incomplete Amounts
Locate the boundaries of the missing words
Segment the Legal Amount Image according to the boundaries
Word recognition
Form Possible Amounts
Comply with syntax
No
Yes
Accept
Eject

Figure 5.14 The recovering process to recognize the incomplete amounts.

5.3.1 Locating the boundaries of the missing words

Connecting characters in handwriting causes character segmentation and recognition errors. However, the whole image of a word might contain these characters and retain sufficient information to identify the word class. If the boundary of these words can be located, the word recognition approach would identify the word classes. These word locations are only the hypothesis, since there is no simple way to tell that the locations derived are the actual boundaries of a word. The connecting characters have various forms. They might be created from two or more characters. In actual handwriting, there are so many ways that two or more characters can be connected but if the connections are limited to be within a word, some patterns of the characters can be analyzed and used to locate the possible word boundary. The connecting characters can be divided into three groups: connection between body characters, between body and high characters, and the combination of the first and second groups.

If the touching pattern is formed by two or more characters in the body group, its width should be larger than the average width of any single character and the input sequence will be reduced. Figure 5.15 shows an example of connecting between the body characters. The amount shown is เจ็ด-ร้อย-เก้า. There is connection between the character [ร], [อ] and [ย] that makes all characters connected as one blob. The normal amount has nine body characters but the connecting amount has only seven body characters.

If the touching is formed between two or more characters in the body and the high group, its height should be higher than the average height of any character given that most of the characters in an amount are not [ส] and [ป] and the input sequence length remains the same. Figure 5.16 shows an example of connecting between the body and high group characters. The amount shown is หนึ่ง-ร้อย-หก. There is a connection at ร้อย

between [ร] and [โ]. The number of body characters of the normal and connected amounts remains the same.

Another type of the connection is the combination of the previous descriptions. Three or more characters can be connected. The image shape will be higher and wider than other single characters. There is a connection between the characters [ร], [โ], [อ] and [ย] which makes all characters connected as one blob. The normal amount has seven body characters but the connecting amount has five body characters. Figure 5.17 shows the connected characters of mixed types.

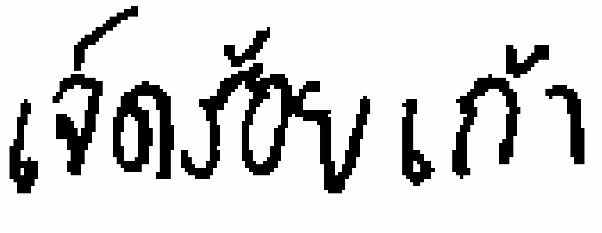


Figure 5.15 The connecting handwriting formed by the body group characters.

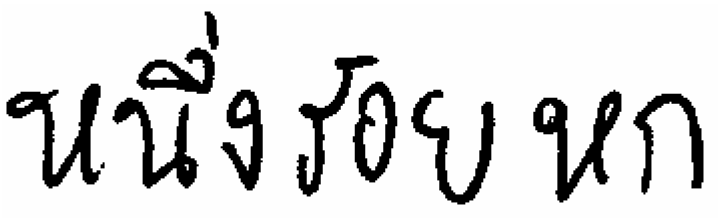


Figure 5.16 The connecting handwriting formed by the body group characters.

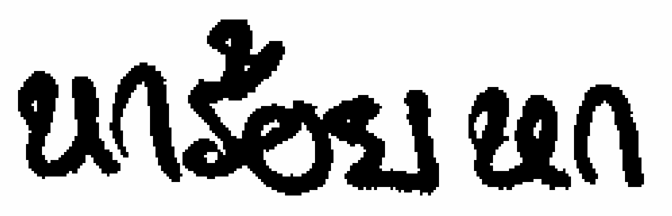


Figure 5.17 The connecting handwriting formed by both body and high group characters.

If the touching characters are only within the words, the boundary of the words can be derived from the locations of these detected irregulars. This approach is possible because the legal amounts have limited lexicon. For the connection between the body character group, the possible combinations of connection are shown in Table 5.4. For example, the connection between characters [พ] and [ั] in the word พัน are one connected component that would be higher than other characters. If this component can be located by its height, the word boundaries can be the positioned before a connected component itself and after the following character.

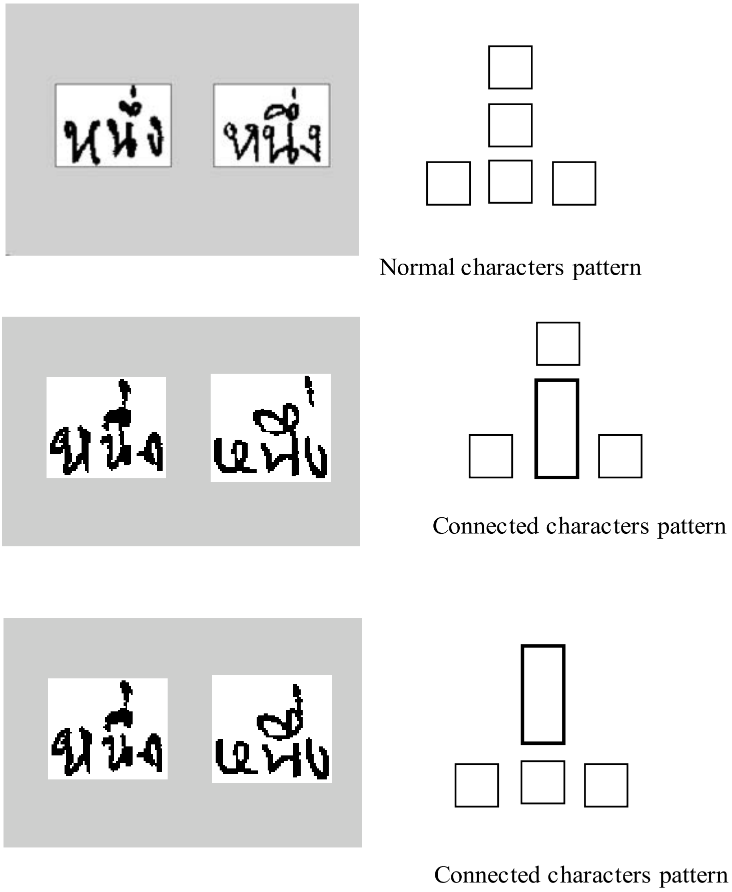
5.3.2 Analysis of the connecting components formed by the body and high characters group

To analyze the connection between the body and high characters group within a word, the amount words are arranged according to the following description. The amount words are arranged into five groups based on character pattern as shown in Table 5.4. In the first group, the word หนึ่ง and หมื่น have three body characters with two high characters above the middle body characters. In the second group, the word เอ็ด เก้า and เจ็ด have three body characters and one high character above the middle body character. The third group members are ล้าน and ร้อย. They have three body characters and one high character above the first body character. The fourth group members are พัน ห้า and สิบ. They have two body characters and one high character above the first body characters.

Group	Words			Character pattern
1	หนึ่ง	หมื่น		□ □ □□□
2	เอ็ด	เก้า	เจ็ด	□ □□□
3	ล้าน	ร้อย		□ □□□
4	พัน	ห้า	สิบ	□ □□
5	สี่	ยี่		□ □ □

Table 5.4 Word grouped by characters patterns.

If the connecting characters exist within a word and between the high and body characters underneath, the example of the connecting character pattern of these words can be shown in Figure 5.18 through Figure 5.22. The square box in these Figures represents one character and the tall box represents the connecting characters.



Normal characters pattern

Connected characters pattern

Connected characters pattern

Figure 5.18 The handwriting samples and characters patterns of normal writing and connected writing of the first group.

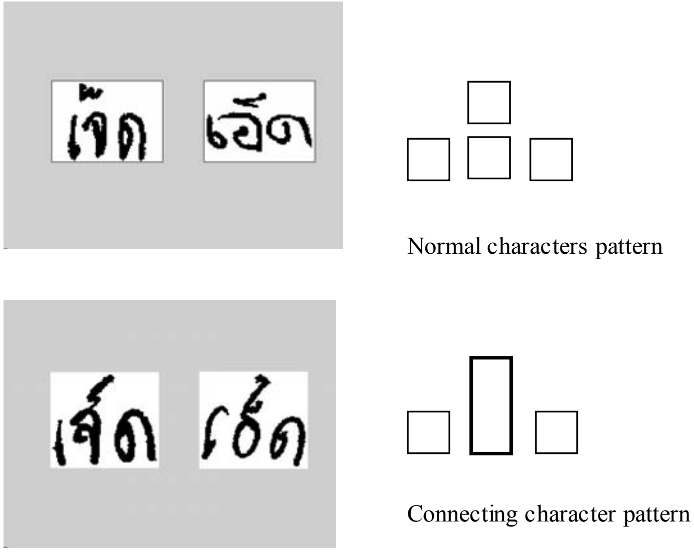


Figure 5.19 The handwriting samples and characters patterns of normal writing and connected writing of the second group.

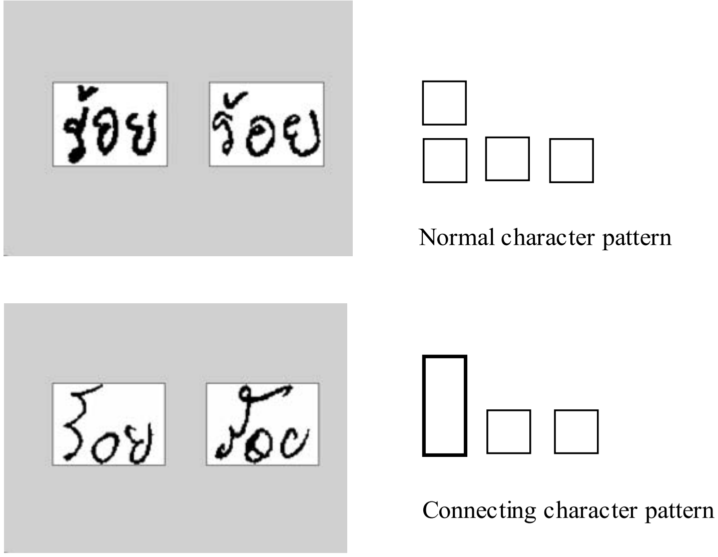


Figure 5.20 The handwriting samples and characters patterns of normal writing and connected writing of the third group.

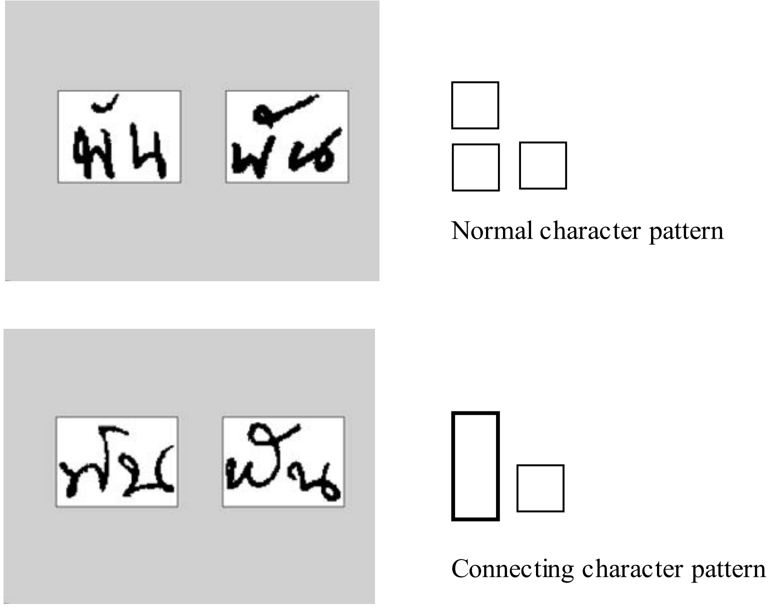


Figure 5.21 The handwriting samples and characters patterns of normal writing and connected writing of the fourth group.

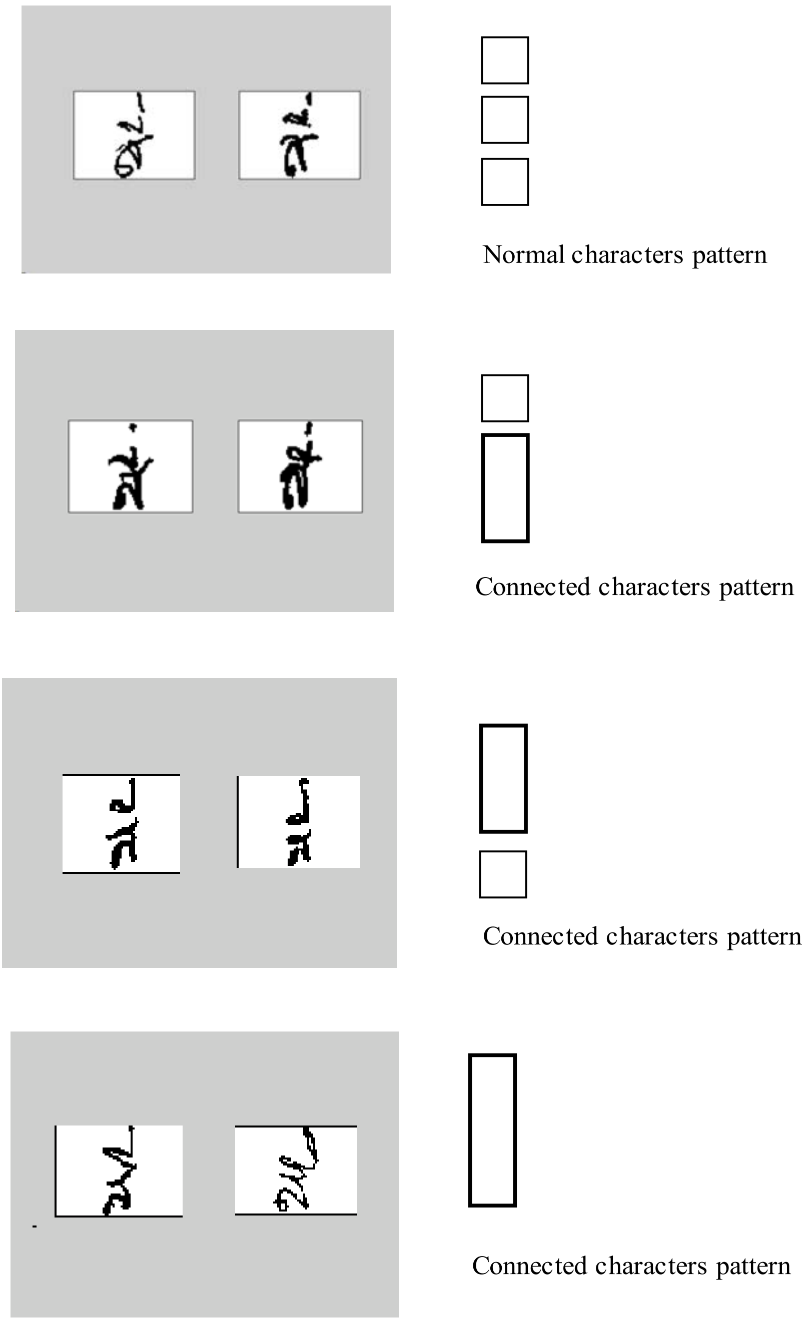

Normal characters pattern
Connected characters pattern
Connected characters pattern
Connected characters pattern

Figure 5.22 The handwriting samples and characters patterns of normal writing and connected writing of the fifth group.

The possible words boundaries of each group can be derived from the position of connecting characters are shown in Table 5.5. There are four types of word boundaries because some patterns are duplicated. The touching characters are formed by body and high group characters. The circle is the position of touching character and the boxes are the nearby characters, the arrow indicates the word boundaries.

Group	Words			Word boundaries
1	**หนึ่ง**	**หมื่น**		□○□ ⟷
2	**เอ็ด**	**เก้า**	**เจ็ด**	□○□ ⟷
3	**ล้าน**	**ร้อย**		○ □□ ⟷
4	**พัน**	**ห้า**	**สิบ**	○□ ⟷
5	**สี่**	**ยี่**		○ ↔

Table 5.5 Word groups and their members and word boundaries.

5.3.3 Analysis of connecting components formed by the body characters group

To analyze the connecting characters between the body characters group, the amount words are arranged according to the following description. The amount words are arranged into three groups based on only number of body character as shown in Table 5.6. In the first group, the word **หนึ่ง**, **หมื่น**, **เอ็ด**, **เก้า**, **เจ็ด**, **ล้าน** and **ร้อย**, all have three body characters. The second group members are **แสน** and **แปด**, both have four body characters. The third group members are **พัน**, **ห้า**, **สิบ** and **หก**. They have two body characters.

Group	Words			Character pattern
1	หนึ่ง	หมื่น		□ □ □□□
	เอ็ด	เก้า	เจ็ด	□ □□□
	ล้าน	ร้อย		□ □□□
2	แสน	แปด		□□□□
3	พัน	ห้า	สิบ	□ □□
	หก			□□

Table 5.6 Word grouped by body characters patterns.

If the connecting exists within a word and between the body characters nearby, the example of the connecting character pattern of these words can be shown in Figure 5.23 through Figure 5.25.

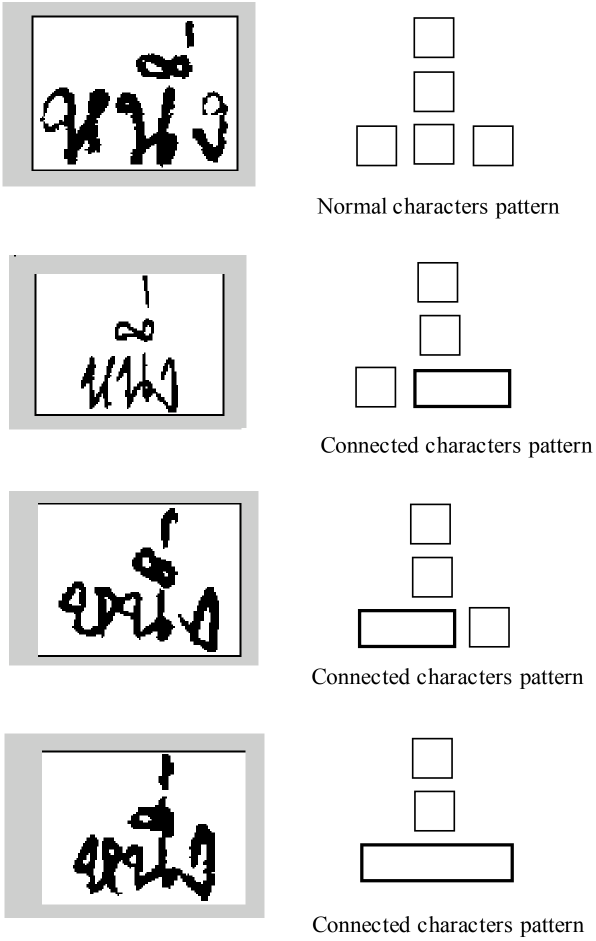

Normal characters pattern
Connected characters pattern
Connected characters pattern
Connected characters pattern

Figure 5.23 Handwriting samples and characters patterns of normal writing and connected writing for word in group I. The connections are formed between the body group characters.

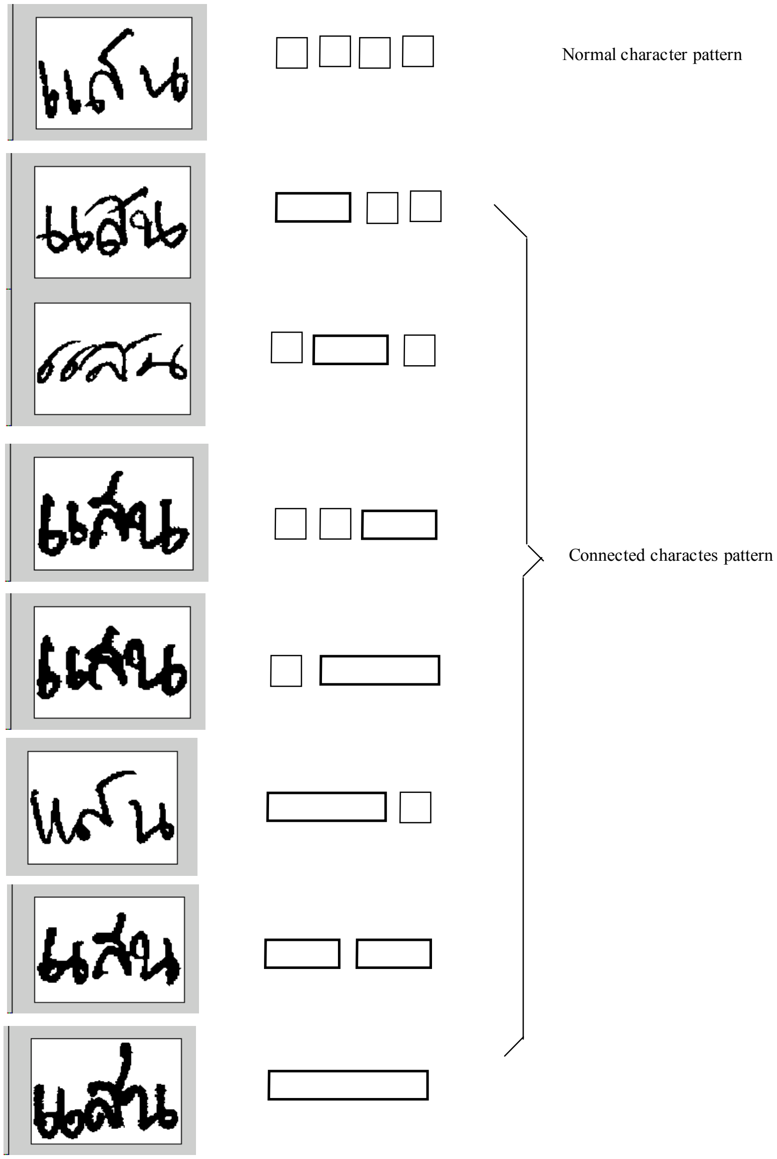


Figure 5.24 Handwriting samples and characters patterns of normal writing and connected writing for word in group II. The connections are formed between the body group characters.

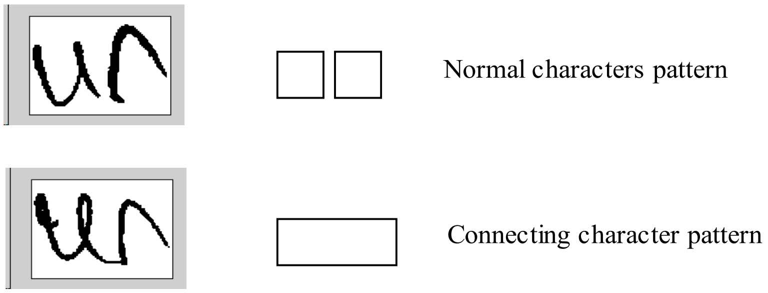


Figure 5.25 Handwriting samples and characters patterns of normal writing and connected writing for word in group III. The connections are formed between the body group characters.

The possible word boundaries of each character groups that can be derived from the position of connecting characters are shown in Table 5.7. There are six types of word boundaries because some patterns are identical. The touching characters are formed between the body group characters. The circle is the position of touching character and the boxes are the nearby characters, the arrow indicates the word boundaries.

Group	Words			Word Boundaries
1	หนึ่ง	หมื่น		○□ ↔ □○ ↔ ○ ↔
	สอง	สาม		
	เอ็ด	เก้า	เจ็ด	
	ล้าน	ร้อย		
2	แสน	แปด		○□□ ↔ □○□ ↔ □□○ ↔ □○ ↔ ○□ ↔ ○ ↔
3	หก			○ ↔
	พัน	ห้า	สิบ	

Table 5.7 Three groups of words and their word boundaries.

5.3.4 Word segmentation

The previous section describes a systematic approach to derive a number of possible word boundaries from the connecting characters. These boundaries are only the hypothesis and need to be verified by word recognition. The boundaries will be used to segment the hypothesis word image. The word segmentation can be straightforward done by cutting the amount image from the position in front of the left boundaries and end at the back of the right boundary. However, there are some cases where the upper characters are irregular, i.e., very long or very big or drawn out of their regular position and overlapping with other words. Figure 5.26 shows the image of a legal amount เก้า-ร้อย-สิบ (910). This image is supposed to be segmented into three word images. The first word is เก้า that has no overlap with other words. The second word is ร้อย where the end part of this word is overlapped with the following word. The third word is สิบ where some part of vowel [ิ] is outside word boundaries.

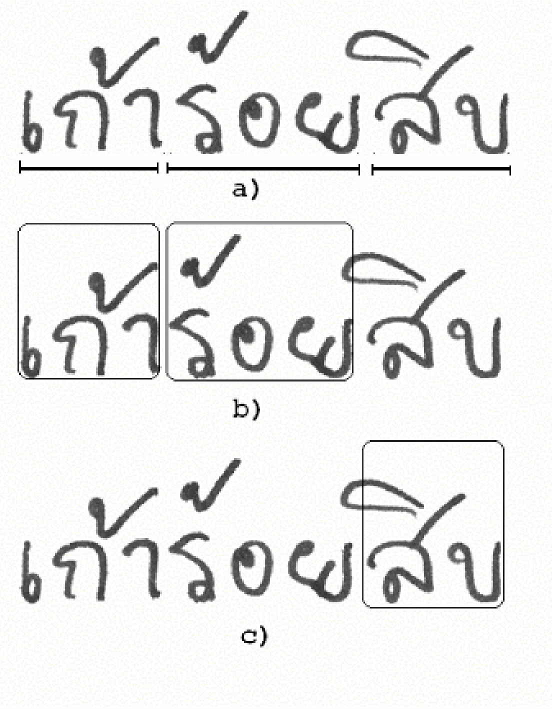


Figure 5.26 The legal amount of เก้า-ร้อย-สิบ (910). (a)The original image and approximated words spaces. (b)Word boundaries of เก้า and ร้อย. (c)Word boundaries of สิบ.

To segment the image of words correctly, a routine is designed to check if any stroke crosses the word boundaries, then whether or not this stroke should be added to the word image or should be ignored. The decision to remove and add the cross boundary objects is based on a proportion of number of the pixels of the object that cross the boundaries. Figure 5.27 shows the word segmentation that includes and excludes overlapping.

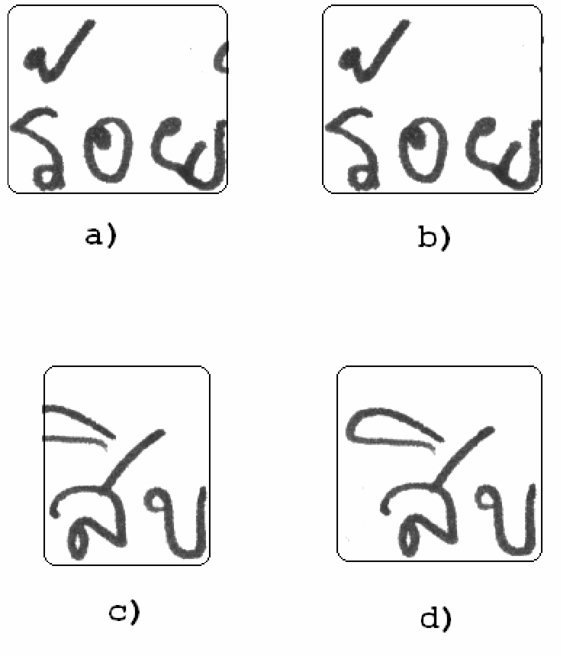


Figure 5.27 Word segmentation. (a)The initial image of word ร้อย. (b) The image of word ร้อย excludes the overlapping part. (c) The initial image of word สิบ. (d) The completed image of word สิบ includes the overlapping part.

5.4 Summary

This chapter has shown techniques to improve the recognition performance in two parts. The first part concerns processes at the preprocessing and recognition process stage by dividing characters into subclass with different recognition systems, combining multiple features, and using the recognition with the ranking results. The second part is concerned with processes at the post processing stage. The system uses information from lexicon and syntax to form the possible answers and rejects those that are not compiled with the syntax rules. In case no amount is accepted, the irregular shape components are searched and the possible boundaries of the word that originate those components are generated. The word segmentation will separate the hypothesis words from the amount image. These images will be identified by the word recognition process. The results will be joined with previous words matched to form possible amounts and check with the syntax again. The information of character and word recognition will be used to score the most sensible decision answer for the legal amount.

Chapter 6

Experiment and Results

This Chapter uses the concepts proposed in Chapter 5 to verify if the techniques can improve the recognition performance and the result analysis. The design and collection of the handwriting data set used in this Chapter are introduced in Chapter 4.

6.1 The writing level separation process

The experiments in this section are conducted to verify if the character recognition rates are improved when the characters are separated into high and body character groups and to evaluate how well the level separation algorithm separates the test data. The data used in the experiments are the Thai handwriting samples of 2500 characters from 25 character classes. These samples are divided into two groups: 800 characters in the high group and 1700 in the body group. The samples are divided into training and testing data with ratio of 3:1. The results show that when the input characters are separated into the body and high group characters, the recognition rates of both groups improve. The average recognition rate for each feature and each character group is summarized in Table 6.1.

Features	Training data set	Testing data set
Cavity image	80.86	46.10
edge direction type I	94.20	73.5
edge direction type II	90.90	64.54
outer contour	93.80	77.30
transition profile	94.90	77.30
Fourier descriptor type I	73.18	44.68
Fourier descriptor type II	80.78	52.48
Fourier descriptor type III	67.22	51.6
Fourier descriptor type IV	92.24	52.48

Table 6.1 Average recognition rate with the body character group in percent.

Features	Training data set	Testing data set
Cavity	95.83	63.64
edge direction type I	97.83	80.83
edge direction type II	97.67	77.27
outer contour	96.50	72.73
transition profile	96.5	75.76
Fourier descriptor type I	87.81	65.15
Fourier descriptor type II	92.0	60.61
Fourier descriptor type III	84.83	56.60
Fourier descriptor type IV	97	54.55

Table 6.2 Average recognition rate with the high character group in percent.

In order to evaluate how well the algorithms perform separating the input characters into the correct group level, the extracted characters from the test amounts are used. The total number of characters are 3,126 characters from 154 legal amounts. The algorithms can separate the characters into correct classes with the correct rate of 99.87 percent. The results are shown in Table 6.3.

	High group	Body group
High group	643 (99.68%)	2 (0.31%)
Body group	2 (0.08%)	2,489 (99.02%)

Table 6.3 The level separation of the characters extracted from the test amounts.

From 2,489 body group characters, 2,281 are isolated characters, 136 are connecting characters formed by body and high group characters, and 72 are connecting characters formed by more than two characters or formed between body characters.

Isolated Characters	Connected components	
	Body-high	Others
2,281 (91.64%)	136 (5.46%)	72 (2.89%)

Table 6.4 The number of single characters and connected components of test amounts of a total of 2,489 components.

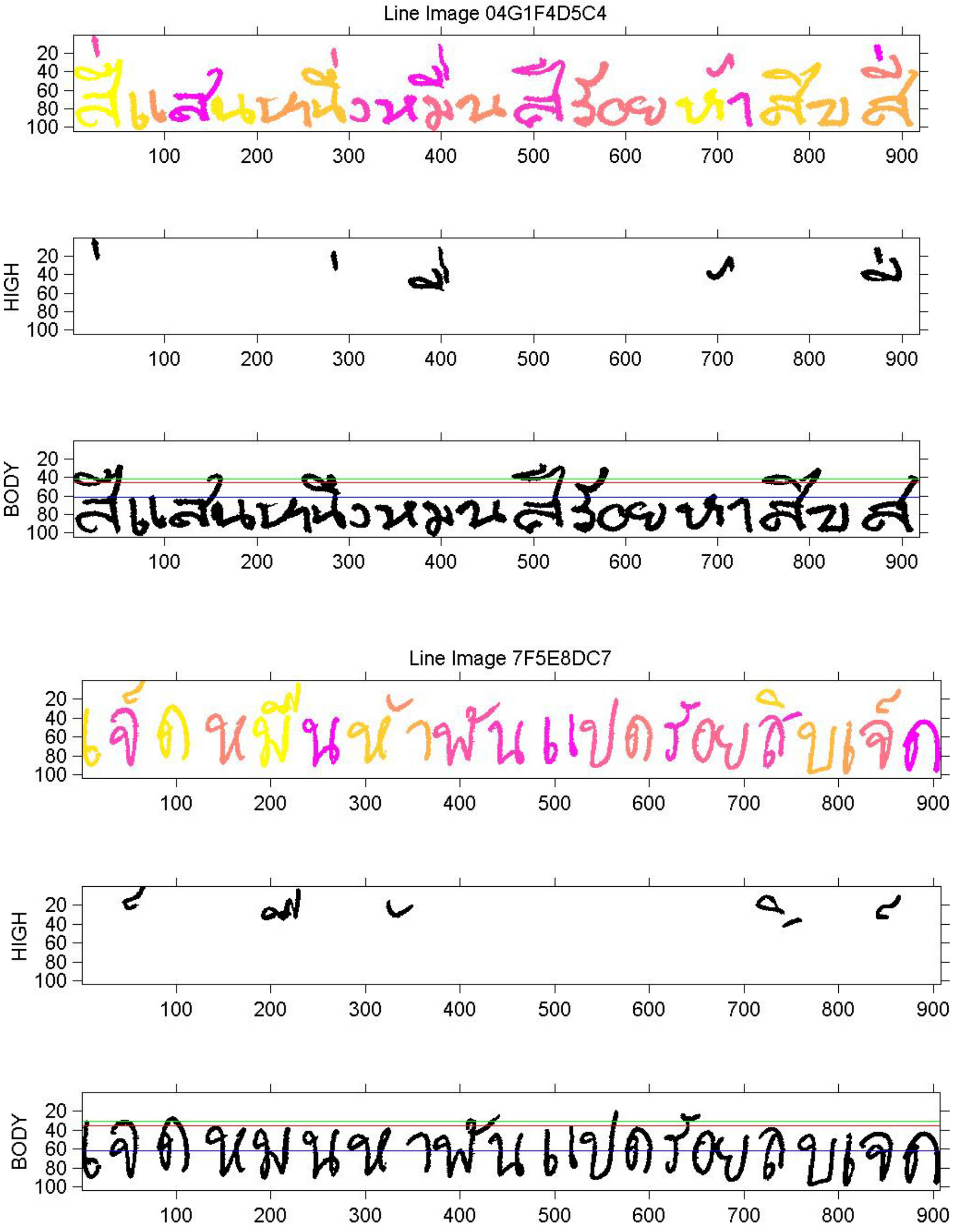


Figure 6.1 Writing level separation results.

6.2 The feature combinations for recognition

The single features used in the preliminary experiments are reconsidered as each of them can effectively capture different classes. Howerver, the combination of these features can detect and classify with better recognition rates. Empirical experiments are employed to seek the best combination of the features.

The features for Thai characters in the experiments include the transition profiles with zoning, the outer contour, the edge direction type I, the edge direction type II, the Fourier descriptor type I through IV and image cavity. However, results from the preliminary experiments indicate that different Fourier descriptors show no significant difference at the captured classes but with different recognition rate. Therefore, the combination experiments will select Fourier descriptor type IV, which has the highest rates among its groups. The combination with image cavity is also not shown here because the results show no improvement in recognition rates. The results of the feature combination indicate that the best combination of the feature is the outer couture, the transition profiles, and the edge direction type I. The recognition rates of each features combination are shown in Table 6.5.

Features	Training set	Test set
outer contour, transition profile	96.8	79.76
outer contour, transition profile and edge direction type I	98.27	82.82
outer contour, transition profile, and edge direction type II	95.92	80.71
outer contour, transition profile, edge direction type I, and Fourier descriptor type IV	96.1	78.82
outer contour, transition profile, edge direction type II and Fourier descriptors type IV	95.14	76.94

Table 6.5 Average recognition rate for each feature for the body group characters in percent.

Features	Training set	Test set
outer contour, transition profile	97.2	78.5
outer contour, transition profile and edge direction type I	95.8	84.5
outer contour, transition profile, and edge direction type II	97.0	83.5
outer contour, transition profile, edge direction type I, and Fourier descriptor type IV	98.5	80.0
outer contour, transition profile, edge direction type II and Fourier descriptors type IV	97.5	79.5

Table 6.6 Average recognition rate for each feature for the high group characters in percent.

	ห	น	ง	ส	อ	า	ม	ก	เ	จ	ด	ป	บ	ย	ร	พ	ล
ห	76	12	-	-	4	-	-	-	-	-	4	-	-	-	4	-	-
น	4	60	-	4	-	-	12	-	-	-	-	4	4	-	-	12	-
ง	-	-	88	-	-	4	-	-	4	-	-	-	-	-	4	-	-
ส	-	-	-	92	-	-	-	-	4	-	-	-	-	-	4	-	-
อ	-	-	-	-	76	-	-	-	-	-	12	-	-	4	4	-	4
า	-	-	-	-	4	84	-	4	8	-	-	-	-	-	-	-	-
ม	4	-	4	-	-	-	76	-	-	-	4	-	8	-	-	4	-
ก	4	-	-	-	-	-	-	88	-	-	4	-	-	-	-	-	4
เ	-	-	4	-	-	4	-	-	88	-	-	-	-	-	4	-	-
จ	-	-	-	-	-	8	-	-	-	88	-	-	-	-	-	-	4
ด	-	-	-	-	-	-	4	-	-	-	84	-	-	-	-	-	12
ป	-	-	4	-	-	-	-	-	4	-	-	92	-	-	-	-	-
บ	-	8	-	-	-	-	-	-	-	-	-	4	88	-	-	-	-
ย	-	8	-	-	4	-	-	-	-	-	-	-	4	84	-	-	-
ร	-	-	8	4	-	4	-	-	-	-	-	8	-	-	76	-	-
พ	4	-	-	4	-	-	4	-	-	-	-	4	-	-	-	84	-
ล	-	-	-	8	-	-	4	-	-	-	4	-	-	-	-	-	84

Table 6.7 Confusion matrix of a recognition system using outer contour, transition profile and edge direction type I features with body group characters. The average recognition rates for the training set and testing set are 98.27 and 82.82 percent respectively.

	ึ	่	ี	้	๊	ิ	ั	ื
ึ	80	-	12	-	-	4	-	4
่	-	100	-	-	-	-	-	-
ี	8	-	68	-	8	4	4	8
้	8	-	4	72	-	-	16	-
๊	-	-	-	-	96	4	-	-
ิ	4	-	-	-	-	92	4	-
ั	-	-	-	20	8	-	72	-
ื	4	-	-	-	-	-	-	96

Table 6.8 Confusion matrix of a recognition system using outer contour, transition profile with zoning, edge direction type I with high characters group. The average recognition rates for the training set and testing set is 95.8 and 84.5 percent respectively.

6.3 The recognition with ranked output

There are two approaches to implement the ranked output recognition system. The first approach is an implementation system using the neural network. A total of 1,700 handwriting samples in the body group are used to train the neural network. The neural network is trained with the target error of 0.02. The best results from twenty experiments are selected. The cumulative top-1 to top-3 recognition rates for the training data set are 98.27, 98.82 and 99.06 percent respectively. This network is then used to recognize the characters extracted from the test amounts. The cumulative top-1 to top-3 recognition rates are 82.82, 89.65 and 92.94 percent respectively.

In the second approach, the recognition system is implemented using the *k*-NN concept. The same handwriting data set are used as the models and stored in the system. Empirical experiments are conducted to find the value of *k* that gives the best results. The highest recognition rate is archived when k is twenty and the cumulative top-1 to top-3 rates are 80.94, 89.71 and 89.71 percent respectively. Therefore, the neural network recognizer will be used in the next experiments. The samples of the ranked recognition from neural network results are shown in Figure 6.2.

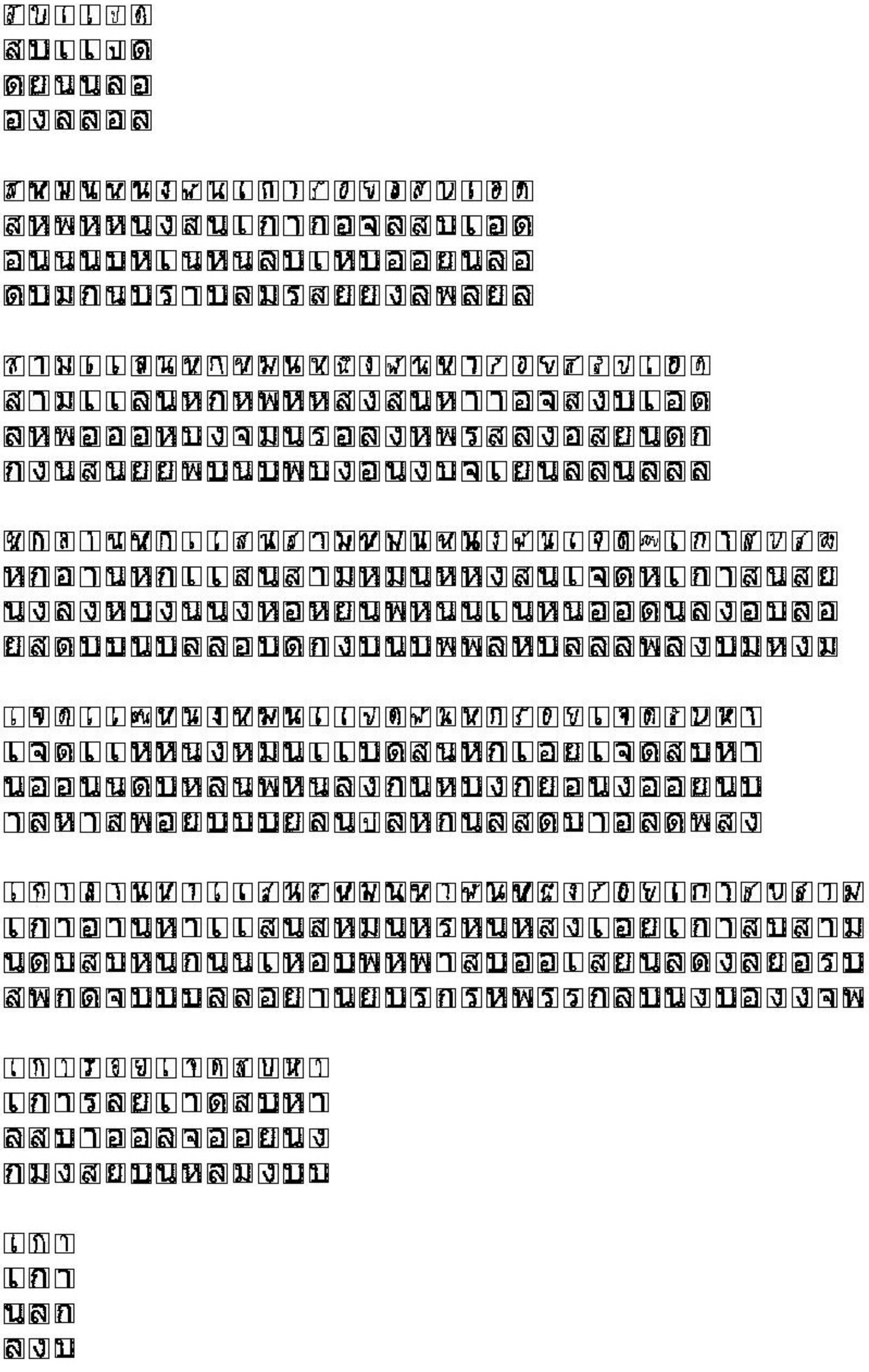


Figure 6.2 Examples of the test legal amount and the ranked recognition output. There are eight amounts in this Figure. The input characters are shown in the first row and the ranked recognition characters are shown in the vertical row.

6.4 The postprocessing process

The ranked character results in section 6.1 are matched with lexicon matching to locate the possible words. These words are joined together to form choices of legal amounts and checked with syntax. Any amounts that are not compiled with the syntax are rejected. There are 47 rejected amounts by the syntax check from a total of 154 amounts. The average rejection rate is 30.52 percent. The recognition of the 107 amounts is 100 percent correct. The rejection rates of each word length are shown in the second row of Table 6.9.

6.5 The recovery process

The rejected amounts are searched for any anomaly components. The boundaries of the hypothesis words are generated with respect to the type of anomaly. The hypothesis words are segmented from the legal amount images and recognized by the word recognition. The word results are then joined together with the matched words described in the previous section to form the legal amounts. These legal amounts are checked whether or not they comply with the syntax again.

To implement the word recognition, a total of 3,500 handwriting samples from 17 legal amount words are used to train the systems. The selected features are the width and height ratio, the transition profiles, the outer contour, and the edge direction type I. The ranked recognition rates of the training set, top-1 to top-3 cumulative are 72.80, 85.77 and 92.44 percent respectively. From the 47 rejected amounts, 32 amounts are detected with anomaly components and 24 amounts are correctly recognized by the word recognition 8 amount are misrecognized and rejected by the syntax check. The average rejection rate is improved to 14.9 percent. The amounts that pass the syntax check are 100 percent correct. Table 6.9 shows the rejection rate after the recovery processes in the third row.

Word length	1	2	3	4	5	6	7	8	9	10	11	12	13
Total amounts	25	9	17	2	21	4	20	7	15	2	21	3	10
Rejected amounts without recovery processes	0 (0%)	5 (55.55%)	3 (11.65%)	0 (0%)	6 (28.6%)	1 (25%)	6 (30%)	1 (14.3%)	7 (46.7%)	1 (50%)	10 (46.17%)	2 (66.7%)	5 (50%)
Rejected amounts with recovery processes	0 (0%)	1 (9.1%)	2 (13.3%)	0 (0%)	1 (4.8%)	0 (0%)	4 (20%)	1 (14.3%)	4 (26.7%)	0 (0%)	4 (19%)	1 (33.3%)	3 (30%)

Table 6.9 Rejection rate without and with recovery processes.

6.6 Summary

This Chapter shows the results of the proposed techniques. Dividing the input characters into *body* and *high* character groups indicates the improvement in recognition rate. The experiments that combine features indicate that the best combination of features are outer contour, transition profiles, and edge direction type I. The writing level separation algorithm performs very well on medium-quality handwriting. The ranked characters recognition and lexicon matching can reveal most of the words in the input characters. The syntax check can reject the unwanted amounts and leave only the correct amounts. The anomaly detection can detect the anomaly components and segment hypothesis words. The ranked word recognition can recognize most of the input. Finally, these words are joined to form amounts and checked against the syntax again. From 154 amounts written by 10 writers, the rejection rate is improved up to 14.9 percent with recovery process. The recognition rate is 100 percents.

Chapter 7

Conclusion and Future Works

7.1 Summary

In this thesis, a recognition system for Thai handwriting legal amount is described as well as its enhanced ability to archive at the high recognition rate. Thai handwriting in legal amounts is a difficult problem and a new field within the handwriting recognition research. This is the first exploration in this area. Many fundamental resources such as the data set did not exist when this study began. The focus of this thesis is on the following issues that bring significance to the implementation of a handwriting recognition system. From 154 amounts written by 10 writers, the rejection rate is 14.9 percent with the recovery processes. The recognition rate for the accepted amount is 100 percent.

- **The design and collecting of Thai handwriting data set**

  The handwriting data set is the fundamental component of the handwriting recognition system. The data sets collected in this study are handwriting characters, words and amounts. The training data are designed and selected to have equal samples per class using repetitive patterns or characters and words, while the test data are designed to represent general amounts. To create the test data, the amounts are divided into seven ranges by the number of digits, and equal samples per range are randomly selected. Then the handwriting data are collected from a number of Thai native volunteers. Preliminary experiments of a basic character recognition system using various features are conducted to show the limitations of the system. The results are then compared with the proposed techniques to improve the system.

- **Improving techniques for the character recognition**

  The character recognition is the basis of the handwriting recognition system. Techniques are proposed at the preprocessing and recognition processing level. The characters are divided into smaller subgroups by their writing level, named *body* and *high* groups. The recognition rates of both groups increase based on their distinguished features. This success is based on how well the algorithms separate characters into the right group. The level group separation algorithms are implemented using the size and position of characters. The single features used in the preliminary experiments are reconsidered as each of them can effectively capture different classes. The combination of these features could detect and classify more data. Empirical experiments are employed to test the best combination of the feature. Traditional recognition systems, which normally give only one answer indicating the highest scored, are modified to give ranked answers ordered by score to cover all of the possible character classes.

- **The postprocessing process**

  The ranked characters results cannot be used directly because each input has *n* possible answers. The postprocessing process uses a prior knowledge such as lexicon and syntax to select and filter only the sensible answers. The lexicon matching algorithms are implemented to match the ranked characters from the body group characters with all legal amount words. The exact match uses all characters in a word to compare with the ranked results. The partial match allows one replacement error in a word when compared with the ranked results. The combination of both matching types reveals possible words along the input sequence. The possible words are joined together to form choices of amounts, named *amount path*. Each path will be checked whether or not they comply with the syntax of the legal amounts.

- **The detection and correction of anomaly characters**

  The anomaly in handwriting caused by touching or broken characters is normally detected by its size and shape. The combination of possible patterns that can be the word boundaries can be pre-defined and used to segment the hypothesis words. The whole word image is immune to the noise of excessive or broken strokes and can be identified by word recognition. The word recognition results are then joined with the matched words and checked by the syntax rules.

7.2 Future works

- **Handwriting data set**

  In this study, a total of 11,655 characters, 3,500 words samples and 154 amounts are collected. This data set could be regarded as preliminary for testing concepts. More samples from a large group of writers are needed to represent realistic handwriting models. For example, NIST's special database 19 (NIST) contains 800,000 binary images of handwriting alphanumeric from 3,600 writers. The samples from real application documents should be considered as in the future design and collecting process.
- **The improvement of characters recognition**

  The characters recognition can be improved by many means, such as finding new features, using multiple features or using multiple recognizers. In this study, the features perform well for the general shapes but no features are designed specifically for Thai characters. The features combination is prepared by adding and selecting the most variant attributes but this is not a guarantee that these combinations are the best for the recognition results. There are some methodologies in feature selections, such as multiple discriminate analysis and genetic algorithm selection that should be performed to find the best solution.

- **The improvement of word recognition**

  The word recognition can be improved by the same means used in of the character recognition, such as finding the better features and the recognizer. The features used for the word image should be used with multiple components and with a high-immunity to the words' shapes.

- **The matching algorithm**

  The partial match algorithm in this study allows one replacement error, which is robust enough for the testing amounts in this study. However, the error condition should be extended to other types such as insertion or deletion and to more than one error per word. In extension to the monetary document reading system, the verification of the result in the final stages in this study with the result from the digit recognition systems would be an essential process.

7.3 Contributions

I believe that the works presented in this thesis have made the following contributions in the field of Thai handwriting legal amount recognition.

- Techniques to improve character recognition rate by dividing the characters class into smaller subclasses, combining features, and using a ranked recognition system.

- A robust lexicon matching algorithms based on information from the body group characters.

- A design of the syntax check algorithm for Thai legal amount.

- The detection of anomaly created from touching and broken characters.

- The design of a word recognition process using the holistic approach.

Bibliography

[1] Airphaiboon, S. and Kondo, S., **Recognition of Handprinted Thai Characters using Loop Structures**, Sep. 1996, IEICE, pp. 1296-1304.

[2] Airphaiboon, S. and Kondo, S., **Offline Handwritten Thai Characters From Word Scripts**, Oct. 1994, Internal Conference In Pattern Recognition: IEEE, pp. 445-449.

[3] Anderson, J. Pellionisz, A. and Rosenfeld, E., **Neural Computing 2: Directions for Research,** MIT Press, Cambridge CA, 1990.

[4] Andrews, H., **Multidimentional rotations in features selection**, IEEE transaction of Computers, Vol.20, pp. 1045-1051, Sep. 1971.

[5] Anigbogu, J. and Belaid,A., **Hidden Markov Models in Text Recognition,** International Journal of Pattern Recognition and Artificial Intelligence, 1995, Vol. 9, No. 6, pp. 925 – 958.

[6] Balaid, Y., Belaid, A., and Turolla, E., **Item Searching in Forms: Application to French Tax Form**, International Conference on Documentation Analysis and Recognition, Montreal, Canada, pp. 744-747, 1995.

[7] Bengio, Y., **Markovian Models for Sequential Data**, Neural Computing Survey, pp. 129-162, 1999.

[8] Bailey, R., and Sirnath, M., **Orthogonal moment features for use with parametric and non-parametric classifiers**, IEEE transaction on Pattern Analysis and Machine Intelligence, Vol.18, 1996, pp. 389-399.

[9] Bunke, H., Roth, M. and Schukat-talamazzini, **Off-line Cursive Handwriting Recognition using Hidden Markov Models**, Pattern Recognition, Vol. 28, No. 9, pp. 1399-1413, 1995.

[10] Burr, D., **Elastic matching of line drawings**, IEEE transaction on Pattern Analysis and Machine Intelligence, Vol. 3, pp. 708-713 Nov, 1981.

[11] Chatwiriya, W., **Printed Thai Character Recognition by Hypothesis and Evaluation**, Master Thesis, King Monkut Institute of Technology Ladkrabang., Bangkok, Thailand 1994 (in Thai).

[12] Choruengwiwat, P., Jitapunkul, S., Wuttissittikulkij L. and Seehapan, P., **Distinctive feature analysis for Thai handwritten character recognition based on modified stroke changing sequence,** Nov. 1998, IEEE Asia-Pacific Conference on Circuits and Systems , pp. 543-546.

[13] Cohen, E., Hull, j., and Srihari, S., **Control Structure for Interpreting Hand-written Address,** IEEE Transaction on Pattern Analysis and Machine Intelligence, 16(10), pp. 1049-1055, 1994.

[14] Chow, C., **An Optimum Character Recognition System Using Decision Functions**, IRE Transaction on Electronic and Computing, EC-6, No. 4, 1957 pp. 247-254.

[15] Divijver, P., and Kittler, J., **Pattern Recognition: a Statistical Approach**, Prentice-Hall Englewood Cliffs, NJ, 1982.

[16] Duda, R. and Hart, P., **Pattern Classification and Scene Analysis**, John Wiley & Sons, Inc., New York, 1979.

[17] deVel O., Wangsuy S., and Coomans, D., **On Thai character recognition**, Nov 1995, IEEE Int. Conf. Neural Network – Proceeding, pp. 2095-2098.

[18] Dodani, S., Breeding, K. and McGhee, R., **Aircraft identification by moment invariants**, IEEE Transaction on Computing, Vol. C-26, pp. 39-45, 1977.

[19] Doug C., **How Do Thais Tell Letters Apart?** 1995, Research Paper 4, Center for Pure and Applied Research in Computational Linguistics, Chulalongkorn University, Bangkok, Thailand.

[20] Doug C., **Fuzzy Letters and Thai Optical Character Recognition**, 1995, Symposium on Natural Language Processing' 95, Kasetsart University, Thailand.

[21] Doug C., **Font Design for Thai/English Typesetting**, 1995, Symposium on Natural Language Processing' 95, Kasetsart University, Thailand.

[22] Doug C., **How to Read Less and Know More: Approximation OCR for Thai**, SIGIR, 1997.

[23] Dougherty, E., **Digital Image Processing Methods**, Marcel Dekker Inc., New York, 1994.

[24] Dugad, R., and Desai, U., **A Tutorial on Hidden Markov Models**, Technical Report: SPANN-96.1, May 1996, Dept. of Electrical Engineering, Indian Institute of Technology, Bombay, India.

[25] Dunn, C., and Wang, P., **Character segmentation techniques for handwritten text-a survey,** 11th IAPR International Conference on Pattern Recognition, 1992. Vol.II. Conference B: Pattern Recognition Methodology and Systems, pp. 577 – 580.

[26] Dudani, S., Breeding, and K., McGhee, **Aircraft identification by moment invariants**, IEEE Transaction on Pattern Analysis and Machine Intelligence, C-26 (Jan 1997), pp. 39-46.

[27] Elms, A., **A connected character recognizer using level building of HMMs**, IEEE 12th IAPR International Conference On Pattern Recognition, Oct, 1994, pp. 439-441.

[28] Faure, C., **Pen Based Human-Computer Interaction Handwriting and Drawing Research: Basic and Applied Issues**, Simner, M., Leedham, C. and Thomassen A., eds., pp. 373-385, 1996.

[29] Freitas, C.O., Yacoubi A., Bortolozzi F. and Sabourin R., **Brazilian Bank Check Handwritten Legal Amount Recognition**, Proceedings 13th Brazilian Symposium on Computer Graphics and Image Processing IEEE Computer Soc., Los Alamitos, CA, USA; 2000, pp. 97-104.

[30] Fu, K., **Syntactic Pattern Recognition and Application**, Prentice-Hall, Englewood Cliffs, 1982.

[31] Glauberman, M., **Character recognition for business machines**, Electronics, Vol. 29, pp. 132-1336, Feb. 1956.

[32] Grandidier, F., Sabourin, R., Yacoubi, A., Gilloux, M., Suen, and C.Y., **Influence of Word Length of Handwriting Recognition**, Proc. 5th International Conference On Document Analysis and Recognition (Bangalore, India, Sep. 1999), pp. 777-780.

[33] Granlund, G., **Fourier preprocessing for hand printed character recognition**, IEEE Trans. Comput., C-21 (Feb 1972), pp. 195-201.

[34] Hiranvanichakorn, P. , Agui,T. ,and Nakajima, M., **Recognition Method of Thai Characters by Using Local Features,** Aug. 1984, Trans. IECE Japan, Vol. E67, pp. 425-432.

[35] Hirai, Y. and Tsukui, Y., **Position independent neuro pattern matching and its application to handwritten numerical character recognition**, Neural Networks, International Joint Conference in Neural Network, Vol.3, 17-21 June 1990, pp. 695-701.

[36] Hor, D., **A Recognition System for Handwritten Thai Numerals**, Master Thesis , Asian Institute of Technology, Bangkok, Thailand, 1985.

[37] Hu, M.K., **Visual pattern recognition by moment invariants**, IRE Transaction on Information Theory, Vol. 8, pp. 179-182, 1962.

[38] Jain A., Duin, R., and Mao, J., **Statistical Pattern Recognition; A Review**, IEEE transactions on Pattern Analysis and Machine Intelligence. Vol.22, No.1, 2000, pp. 4-37.

[39] Jain, **Fundamental of Digital Image Processing**, Prentice Hall, 1989

[40] Karoonboonyanan T., **Standardization and Implementations of Thai Language,** Seminar on Enhancement of the International Standardization Activities in Asia Pacific Region **(AHTS-1)**, Japan, March 1999.

[41] Kennr, S., Anisimov, V., Baret,O., Gorski,N., Price, D., and Simon, J., **The A2iA Check Reader: A Family of Bank Check Recognition Systems**, Proceedings of 5 th International Conference on Document Analysis and Recognition, Vol. 1, Bangalore, 1999, pp. 523-526.

[42] Kennr, S., Anisimov, V., Baret, O., Gorski,N., Price, D., and Simon, J., **A New A2iA Bank Check Recognition System in Automatic Bankcheck Processing**, edited by S. Impedovo, P.S.P. Wang and H.Bunke, , World Scientific Publication, Singapore, 1997, pp 43-86.

[43] Khunasaraphan, C. and Lursinsap, C., **44 Thai handwritten alphabets recognition by simulated light sensitivity model,** Nov. 1993, Proceeding of the Artificial Neural Networks in Engineering, pp. 303-308.

[44] Kijsirikul B., Sinthupinyo S., and Supanwansa A., **Thai Printed Character Recognition by Combining Inductive Logic Programming with Back Propagation Neural Network**, 1998, Nov 1998, IEEE Asia-Pacific Conference on Circuits and Systems – Proceeding, pp. 539-542.

[45] Kim, W., and Yuan, P., **A Practical Pattern Recognition System for Translation, Scale and Rotation Invariance**, CVPR'94, Seattle, Washington, June, 1994.

[46] Kimpan, C., Itoh, A, and Kawanishi, K., **Recognition of Printed Thai Characters using a Matching Method,** Nov 1983, Proceeding of. IEE, No.6, pp. 183-188.

[47] Kimpan, C., **Printed Thai Character Recognition using topological Properties Method**, INT. J. Electronics, Vol. 60, No. 3, 1986, pp. 303-329.

[48] Kimpan, C., and Walairacht, S., **Thai Characater Recognition**, Proceeding of the Symposium on Natural language Processing in Thailand, March 1993, pp. 123-166.

[49] Kornai, A., Mohiuddin K., and Connell S., **An HMM-Based Legal Amonunt Field OCR System for checks**, 1995, IEEE Proc. Systems, Man and Cybernetics, pp. 2800-2805.

[50] Kundu, A, He, Y., and Bahl, P., **Recognition of Handwritten Word: First and Second Order Hidden Markov Model Based Approach**, Pattern Recognition, 22, 1989, pp. 283-297.

[51] Kuhl, F., and Giardina, C., **Elliptic Fourier feature of a close contour,** Computer Graphics and Image Processing, 18, pp. 236-258, 1982.

[52] Jain, A., **Artificial neural networks: a tutorial,** IEEE computer, pp. 31-44, March, 1996.

[53] Jain, A., Duin, R., and Mao, J., **Statistic Pattern Recognition: A Review**, IEEE transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No.1, 2000.

[54] Lin, C. and Hwang C., **New form of shape invariants from elliptic Fourier descriptors**, Pattern Recognition, Vol. 20, No.5, pp. 535-545, 1987.

[55] Li, X., Parizeau, M., and Plamondon, R., **Training Hidden Markov Models with Multiple Observations – A Combinatorial Method**, IEEE transactions on Pattern Analysis and Machine Intelligence. V.22, No.4, 2000, pp. 371-377.

[56] Liang, S., Ahmadi, M., and Shridhard, M., **Segmentation of touching Characters in printed document recognition**, Proceedings of the Second International Conference on Document Analysis and Recognition, 1993, pp. 569-572, Oct. 1993.

[57] Liu, Y., and Srihari, S., **Document Image Binarization Based on Texture Features**, IEEE transactions on Pattern Analysis and Machine Intelligence. V.19, No.5, 1997, pp. 540-544.

[58] Lohakan, M., Airphaiboo, S., and Sanworasil, M., **Single Character Segmentation for Handprinted Thai Word**, Fifth International Conference on Document Analysis and Recognition 20 - 22 September, 1999 Bangalore, India.

[59] Lursinsap, C., and Khunasaraphan, C., **Simulated Light Sensitive Model for Handwritten Digit Recognition**, Intetnational Join Conference On Neural Networks 1992, IEEE, New York, Vol. 4, pp. 13-18.

[60] Microsoft Corporation., **Microsoft Windows Codepage: 874 (Thai),** http://www.microsoft.com/globaldev/reference/sbcs/874.htm.

[61] Mohamad, M., and Gader, P., **Handwritten Word Recognition Using Segmentation-Free Hidden Markov Modeling and Segmentation-Based Dynamic Programming Techniques**, IEEE transactions on Pattern Analysis and Machine Intelligence, Vol.18, No.5, 1996, pp. 548-554.

[62] Mohamad, M., and Gader, P., **Generalized Hidden Markov Models-Part II: Application to Handwritten Word Recognition**, IEEE transactions on Fuzzy Systems, Vol. 8, No.1,Feb. 2000.

[63] Mori, S., Suen, C., and Yamamoto, K., **Historical Review of OCR Research and Development**, Proceeding of IEEE, Vol. 80, pp. 1162-1180, 1992.

[64] Nantana R., **The Thai Writing System,** http://thaiarc.tu.ac.Th/host/thaiarc/thai/thaiwrt.htm 1997.

[65] Nagy G., **Twenty Years of Document Image Analysis in PAMI**, IEEE transactions on Pattern Analysis and Machine Intelligence., Vol.22, No.1, 2000, pp. 38-62.

[66] Navikamoon, A., "**Kae Roye Kor Kai**", Sarakadee Publishing, BKK, Thailand, 1993 (in Thai).

[67] Niblack, W., **An Introduction to Digital Image Processing**, pp. 115-116, Printice-Hall, 1986.

[68] National Institute of Standards and Technology (NIST), - http://www.nist.gov/srd/nistsd19.htm, [access Sep 20, 2002].

[69] Pavlidis, T., **Structural Pattern Recognition**, Springer-Verlag, New York, 1977.

[70] Perlovsky, L., **Conundrum of combinatorial complexity**, IEEE Transaction of Pattern Analysis and Machine Intelligence, Vol.20, No.6, pp. 660-666, 1998.

[71] Edberg, P., Gonzalez, J., and Jenkins, J., **Map (external version) from Mac OS Thai character set to Unicode 2.0**. ftp://ftp.unicode.org/Public/MAPPINGS/VENDORS/APPLE/THAI.TXT.

[72] Phokharatkul, P., and Kimpan, C., **Recognition of handprinted Thai characters using the cavity features of character based on neural network,** Nov. 1998, IEEE Asia-Pacific Conference on Circuits and Systems - Proceeding, pp. 149-152.

[73] Plamondon R., and Srihari S., **Online and Off-line Handwriting Recognition: A comprehensive Survey**, IEEE transactions on Pattern Analysis and Machine Intelligence, Vol.22, No.1, 2000.

[74] Pratt, W., **Digital Image Processing**, NY, John Willey & Sons, second ed., 1991.

[75] Rabiner, L. **A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition**, Proceeding of IEEE, 1989, pp. 905-915.

[76] Ratree, S., **Printed Thai Character Recognition**, 1985, Master Thesis, King Mongkut Institute of Technology, Lardkrabang, Thailand.

[77] Resis, T., **The revised fundamental theorem of moment invariants**, IEEE transaction transactions on Pattern Analysis and Machine Intelligence., Vol.13, 1991, pp. 830-834.

[78] Schalkoff, R., **Digital Image Processing and Computer Vision**, 1989, John Willey & Son, NY.

[79] Senior, A., **Off-line Handwriting Recognition: A Review and Experiments**, Technical report TR 105, Cambridge University, Engineering Department, Cambridge, UK.

[80] Shen, D. and Ip, H., **Discriminative wavelet shape descriptors for recognition of 2-D patterns**, Pattern Recognition, 32, 1999, pp. 151-156.

[81] Simon, J., Barat, O., and Gorski, N., **A System for the Recognition of Handwritten Literal Amount of Checks**, International Conference On Document Analysis System, Kaiserslauterm, Germany, pp. 135-155, 1994.

[82] Sirhari, S., **From Pixel to Paragraph: the Use of Contextual Models in Text Recognition**, International Conference on Document Analysis and Recogniton, Japan, pp. 416-423, 1993.

[83] Srihari, S. and Keubert, E., **Integration of Handwritten Address Interpretation Technology into the United State Postal Service Remote Computer Reader System**, Proceeding of the Fourth International Conference Document Analysis and Recognition, Vol. 2, pp. 892-896.

[84] Srihari, S., Shin, Y., Ramanaprasad, V., and Lee, D., **Name and Address Block Reader**, Proceeding of IEEE (84) 7, pp. 1038-1049, Jul. 1996.

[85] Sornlertlamvanich V., Potipiti, T., Wutiwiwatchai, and C., Mittrapiyanuruk, P. **The State of the Art in Thai Language Processing**, Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, 2000.

[86] Tanprasert C., and Koanantakool T., **Thai OCR: A Neural network application,** Nov. 1996, Preceeding of IEEE Region 10 Conference, Vol.1, pp. 90-95.

[87] Tanprasert C., and Sae-Tang S., **Thai Type Style Recognition**, 1999, IEEE International Symposium on Circuits and Systems, May 30 - June 2, 1999, Florida, USA., Vol. IV, pp. 336-339.

[88] Teague, M., **Image analysis via the general theory of moments**, Journal of Optical Society America. Vol. 70, pp. 920-930, 1980.

[89] Teh, C., and Chin, R., **On Image analysis by the methods of moments**, IEEE transaction on Pattern Analysis and Machine Intelligent, Vol. 23, 1990, pp. 1089-1101.

[90] The Thailand Royal Institute, **Thai Character Structure Standard,** Bangkok, 2000. Arun Printing, (in Thai).

[91] Thumwarin, P., and Chittayasothorn, S., **Object-oriented expert system for Thai character recognition,** Nov. 1998, IEEE Asia-Pacific Conference on Circuits and Systems, pp. 153-156.

[92] Thomas J. Hudak. **Some Historical Background of Thai Language,** http://thaiarc.tu.ac.th/thai/thai.htm.

[93] Trier, O., Jain, A., and Taxt, T., **Feature Extraction Methods for Character Recognition – A survey**, Patter Recognition, Vol. 29, No.4, pp. 641-662, 1996.

[94] Vinciarelli A., **A Survey on Off-Line Cursive Script Recognition**, 2000, IDIAP research report IDIAP-RR-43, Dalle Molle Institute for Perceptual Artificial Intelligence, Martigny, Valais, Switzerland, http://old-www.idiap.ch/publications/vincia-00b.bib.abs.html.

[95] Wang, W., Brakensiek, A. Kosmala, A., and Rigoll, G., **HMM based High Accuracy Off-line Cursive Handwriting Recognition By a Baseline Detection Error Tolerant Feature Extraction Approach,** Proc. The Seventh International Workshop on Frontiers in Handwriting Recognition, Sep.11-13 2000. Amsterdam, pp. 209-218.

[96] Wang, S., Chen, P. Lin, W., **Invariant Pattern Recognition by Moment Fourier Descriptor**, Pattern Recognition, Vol. 27, 1994, pp. 1735-1742.

[97] Wakahara, T., **Toward robust handwritten character recognition**, Pattern Recognition Letters, Vol. 14, 1993, pp. 345-354.

[98] Wakahara, T., **Shape matching using LAT and its application to handwritten numeral recognition**, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 16, 1994, pp. 618-629.

[99] Yu, K. and Jain, A., **A Form Dropout System**, Technical report MSU-CPS-96-62, http://www.cse.msu.edu/cgi-user/web/tech/document?ID=295, Aug. 1996.

[100] Zahn, C., and Roskies, R., **Fourier Descriptor for Plane Closed Curves**, IEEE Transaction on Computing, C-21, March 1972, pp. 269-281.