

Printed Chinese Character Recognition

A these presented in partial fulfillment of the requirements for degree of Honors of Science in Computer Science At Massey University, Auckland, New Zealand

Yuan Liu
ID:02170027

2009

Abstract

Chinese character recognition is complicated and challenging because the nature of Chinese characters.

In this paper, a two level classification Chinese character recognition architecture has been studied.

Movement invariants short comings in terms of insensitivity for Chinese OCR and solution have been discussed. A new proposed optimum grid method which can potentially increase overall OCR performance while keeping robust nature of classifiers is proposed. Design of generic multilayer character recognition architecture has also been discussed.

Keywords: OCR;Chinese Charactor Recogonition;Movement Invariant;optimum grid.

Acknowledgements

I would like to thank my supervisor Dr. Andre Barczak for his support, guidance and all the invaluable feedbacks during the tenure of the project.

This project is also based on all computer science knowledge from undergraduate studies and I would like to thank Dr. Napoleon Reyes for his support and effort to make training procedure in this project possible. Thanks to Prof. Ken Hawick and Dr. Chris Scoggings to make study Honors program in Massey University possible and their invaluable teaching. Thanks to Dr. Martin Jonson, Dr. Ian Bond and Dr. Chris Messom that helped polished my knowledge structure and understanding in Computer Science.

I would also like to thank all my lecturers for their support and invaluable teaching during my studies at Massey University.

Table of Contents

1	Introduction	1
2	Literature Review	3
3	Methodology.	9
3.1	Image acquisition and database.	9
3.2	Image enhancements	11
3.3	First level classifier	11
3.4	Partial features finding	12
4	Discussion and implementation.	13
4.1	Pre-processing.	13
4.1.1	Segmentation	13
4.1.2	Binaries.	15
4.1.3	Noise reduction.	16
4.1.4	Thinning.	16
4.1.5	Normalization.	17
4.2	Processing	18
4.2.1	The challenge.	18
4.2.2	Moment invariants classifier.	19
4.2.3	Two level classifier architecture.	22
4.3	Neural networks.	26
4.4	Post processing	29

5 Results and analysis.	30
6 Conclusion.	33
6.1 Summary	33
6.2 Limitations	34
6.3 Feature work	34
7 Reference	35
Appendix	37

1.Introduction

To able computers recognize characters, speak and listen human languages, communicate with human language is a dream in computer industry. By the hard works have been done so far, this dream is becoming true. Character recognition can save human huge amount time typing messages into computers. However, recognize Chinese characters on computer is more challenging than other language scripts because of huge amount character set (Chinese national standard includes 27484 characters); big variation on structures of each individual characters, which is very different than most western countries who uses small amount of symbols; printed Chinese characters have many fonts, sizes and they can be ordered by rows or by columns; there could be complex tables and images combined with texts in printed Chinese file; these above makes automatic printed Chinese characters recognition very difficult. The outcome of the project will be useful for other projects in object recognition area and the implementation can be used and speed up the process in file management systems which require a lot of input such as library management, bank bill management, ID management which make this project important and worth doing.

In 1966, Casey and Nagy from IBM published the first article about Chinese character recognition. In China, the study and research started in late 1970's, in 1980's, Tsinghua University started research on Chinese character recognition, early work was limited on small amount and uniformed Chinese character recognition problems. X.Ding and her team developed a working solution for Chinese character recognition with accuracy for printed Chinese characters 99.5% under Chinese government support under project "863", she is the leading scientist in Chinese character recognition and published more than 80,000 paper. A lot of her and her team's thesis are included in this report. With ten years hard work, Chinese character recognition can be applied on actual texts, computers

can recognize pictures and characters including multi-language, multi-font, complex layout input sources.

Chinese character is the only graphic character used in ideography in the world. Each of them has unique 2D shape representation. Although there are many works have been done relating to this field, they are either not robust enough (Umeda 1978 and Akiyama 1990) or require lots dimensions. Due to the large amount of characters and structure nature of Chinese characters, finding a generic robust algorithm with good accuracy is desired. Eleven invariant moments (J.Flusser, 1993) is used as a classifier function in this report combines with optimum grid for more computational expensive processing.

Chinese character recognition research involves segmentation, thinning algorithms, feature extraction and neural networks. This report will describe Chinese character recognition procedures with focus on feature extraction.

2. Literature review:

Pattern recognition is a wide open area; it includes face detection, pattern recognition, and character recognition. Character recognition can include online and offline character recognition. The main difference is online character recognition records motions of the strokes while offline character recognition is purely based on geometry shape of the character. According to Ding's work, methods are used in offline character recognition can be applied to online recognition but not vice-versa.

Input images for OCR process are normally from scanners and often contains picture and characters (X.Ding,2002), the characters are often in different sizes and fonts. In order to achieve good recognition result, character extraction in preprocessing is essential for the whole recognition process. Preprocessing should split characters from input; eliminate noise; normalize font size and position in process kernel, extract skeleton strokes (thinning).

As feature based methods extracts the features of the character, as features are mathematically computed, it is suitable for computer to solve problem. Thus optical character recognition (OCR) is not truly based on optical anymore; it is more related to mathematical computations (X.Ding,2002). The challenge becomes finding the best feature extraction algorithm and how to extract features.

As mentioned above, each Chinese character has its unique geometry shape and they are formed by strokes, (X.Zhu et al, 1986) recognize Chinese characters can be based on stroke and structural features. For a long time, this method influenced the Chinese character recognition's research direction. However, it is very difficult in practice because relationship between strokes and structures are very unstable. As a result, although it is possible to get some experimental results,

or even some demonstration systems, this kind of method does not cope with commercial products requirements which need to deal with noise, variations of fonts. Simply rely on strokes and structural features cannot efficiently extract features or is practical.

Human visual system is very robust character recognition system which can cope with various shape changes and all sorts of noise in practice. Human eyes can recognize all sorts of characters and images and eliminates noises, it is obvious to study and simulate human visual system will not only benefit character recognition but also image recognition.

It is also need to note that writing and reading are different procedures. Human beings need to write down every stroke. However when reading, it is not necessary to analyze every stroke or structure of each character. The reading process is reading the whole structure and analyzes whole character (X.Ding,2003). Simulate human vision and process characters as a whole provides robust feature in character recognition and provides a solid foundation for implementing in practice.

Human vision system recognition procedure is firstly eyes capture raw images and then images get processed by brain. In another words, vision system is related with previous similar memory. Vision system stores large amount of samples and memory explains the meaning of samples (R.Amheim,1969).

Feature selection determines character recognition best possible performance, on the other hand, classifier does the actual classification, and it determines the actual performance. Thus to get best recognition results, researches are focused on how to get features and how to classify them (X.Ding, 2003).

There are plenty ways to extract features, for example it can be based on edge detection, transform, grid features, key point features, vector line features, these

should be chosen based on actual performance in practice and these contribute to Chinese character research history.

Fixed sized meshes (grid) (Umeda 1978 and Akiyama 1990) algorithms are used to apply sub windows to split character and extract detailed features. This is a simulation on how human write Chinese characters, and it works in some cases. However, efficiency is one of the problems in this kind of algorithm because structure of Chinese character varies a lot, each sub window will not process same amount of valid pixels and work load difference among sub windows are big, this is unlike English character recognition where each character does not contain many strokes. Another problem with fix grid is it is very sensitive to transformation of characters in kernel.

To cope with first problem above, (L.Jin et al,1997)dynamic grids can be applied to base on geometry characters of the character such as density of strokes in certain areas. However, all of them uses grid to calculate features directly on the whole character, a normal 8x8 grid will generates 64 dimensions, which is not efficient.

Follow up recent works including combining elastic meshes and directional line element feature (L.Yang et al,2008) and increased accuracy for offline hand writing character recognition by 1.4% compared with traditional elastic meshes. However, none of grid based algorithms have solution to the sensitivity of transformations nature of grid system.

In this report, optimum grid will be used (Y.Liu, 2009), optimum grid simulates human vision system which human has a tend to focus on small area of image (A. Barczak, 2009), it will make the second level classifier focus on interested areas only and normally generate maximum 3 different interested areas for each character to extract detailed features using other feature extraction methods. As data to be processed are trimmed rapidly, it can afford to perform more

computational expensive algorithm to extract features in order to achieve a better overall speed and accuracy. Optimum grid should be big enough to handle transformations in order to keep robust nature of feature extract algorithm (if there is any), and it should be possible to achieve potential super-linear speed up can be achieved if interested areas are kept small as small as possible.

Haar-like features (Papageorgiou et al,1998) was originally used on face detection, it is also included in OpenCV library. In fact, OpenCV uses Haar-like features to detect faces. As character recognition is strongly connected to face detection, it is possible to apply Haar-like features to character recognition. However, Haar-like features are sensitive to transformations. Although it is possible to (2007 A.Barczak) make Haar-like features invariant to scaling by normalizing kernel size, rotational and shifting transformations are very common in practice as input paper cannot be guaranteed to be placed perfectly and noises always present.

Seven Hu features (Hu,1962) are calculated based on whole image which matches our goal to simulate human vision system. These features are invariant to rotation, transformation and scaling which provides robust feature we needed. However, this has not been widely used in Chinese character recognition yet because some of Chinese characters are rotational sensitive (甲 and 由 are different characters in Chinese and have different meanings). The following up study finds out these seven features are not truly independent. Flusser in 2000 proposed there are eleven invariants up to the fourth order. Movement invariants are used in character recognitions but it does not perform very well in Chinese character recognition area, apart from rotational sensitive characters mentioned above, there are characters in very similar shape (己,巳,巳).

We can conclude that feature extraction computations based on whole image is the trend to recognize Chinese characters but purely based on this will eliminate

distance between similar shaped characters and thus lead to failure of whole system. We need to simulate human vision better, thus the reason to introduce interest area. As a result, first level classifiers (sorter) needs to be applied to sort certain characters into same sub group according to a rule and second level classifier performs further feature extractions on partial characters to generate final output (X.Zhu et al 1987).

As the amount of Chinese characters are huge, it is common that recognition system are limited by processing performance and memory space, simple classifiers are tended to be used (X.DING,2002).

Other related areas include segmentation algorithms, automatic correction and neural networks. Segmentation usually contains two procedures, firstly analyze whole document and assign different area properties (text area, picture area, table area ect.) and split them. Second step is segmenting every single character correctly from text, then it is possible to perform character recognition (X.Ding, 2003). Once there is an error in segmentation procedure, it will cause un-fixable errors in recognition process. In low quality printed texts, errors by false segmentation are 50% of all errors.

Analyze document is another challenging problem as this requires computers can automatically recognize pictures, texts and every character in input image. As layout are very different from one to one, mixture of different type of characters (Chinese character, English character, numbers) are randomly placed in text, above makes it is very difficult to automatically analyze document.

X.Ding also proposed that it is possible to “segment as we go” by combining segmentation information including projection and edge detection. She also considers different segmentation possibilities (X.Ding,2002).

Once we have all the methods to extract features, we need to get final output. It is possible to setup different recognition systems and compare output, use the

most popular output as final output [4][5][6]. However this will require extra computing resource and accuracy can be improved further. Feed forward back propagation neural network shows great potential in area of voice recognition, artificial control, signal processing, solving non-linear problems. Any closure continuous function can be approximated by feed forward back propagation neural networks with one hidden layer (R.Hechi-Nilson,1989) thus any three layer back propagation neural network can match projections on N dimension to M dimension.

Once recognition system generates output, it is possible that some characters are misclassified but is fixable. Input text has meanings and it is connected by nearby characters(X.Ding et al 2003). Such connection can be checked by meaning of word, sentence and syntax; it is also possible to be described by statistical language models. Because such model is based on statistics and information, they are from huge language database, thus it can be used in all kinds of practical nature language processing. In character recognition, using this model to correct errors after processing procedure, can improve overall recognition performance.

3. Methodology

This section will focus on methods used in this report; implementation details and challenges encountered will be discussed in later sections. Results obtained in this report are using C programming language with OpenCV which is a cross platform open source library on an Intel core 2 duo 2.5 GHz machine under Fedora 11 Linux operating system. Proposed system overflow is shown as blow.

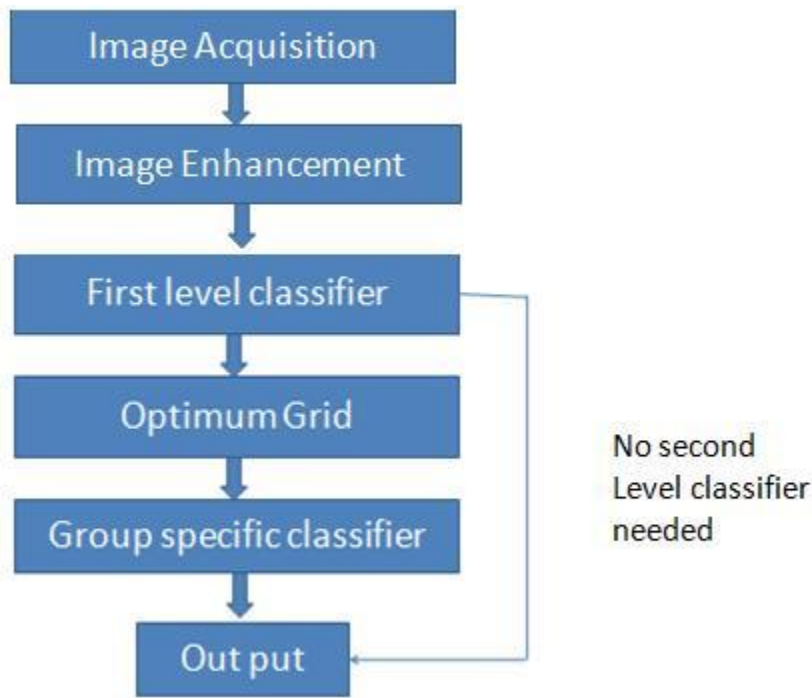


Figure 3.0 proposed work flow for OCR process

3.1 Image acquisition and database

The original input images (documents) are typed under manually under Microsoft Word in font size 36 and then screen captured with Windows built in screen shot tool and saved as .BMP files. Then manually cut each character in .BMP files with same kernel size 96*96 pixels using Gimp. Above procedure is a simulation on segmentation and normalize however this will only generate very small amount of noises compared to practical problems. The goal of this project is to recognize 50 Chinese characters, in order to train neural network and test algorithms, each

character has eleven different rotations and five different fonts. Thus the total number of sample images is 2,550. Screenshots of 50 characters in 5 different fonts are shown as below:

儿八人入大太木十术米口日曰
月目田甲由己巳巳品上下卞卡
志恣一二三五王丸尤妩媚你
尔体本笨笨白勺的女子好

Figure 3.1.1 characters in font simsum

儿八人入大太木十术米口日曰
月目田甲由己巳巳品上下卞卡
志恣一二三五王丸尤妩媚你
尔体本笨笨白勺的女子好

Figure 3.1.2 characters in font kai

儿八人入大太木十术米口日曰
月目田甲由己巳巳品上下卞卡
志恣一二三五王丸尤妩媚你
尔体本笨笨白勺的女子好

Figure 3.1.3 characters in font hei

儿八人入大太木十术米口日日
月目田甲由己巳巳品上下卞卡
志恣一二三五王丸丸尤妩媚你
尔体本笨笨白勺的女子好

Figure 3.1.4 characters in font lishu

儿八人入大太木十术米口日日
月目田甲由己巳巳品上下卞卡
志恣一二三五王丸丸尤妩媚你
尔体本笨笨白勺的女子好

Figure 3.1.5 characters in font MS hei

3.2 Image enhancements

Simple thresh hold method is used to convert grayscale images to binary image in order to reduce calculation complexity. Due to the nature of feature extraction algorithms, it is also needed to reverse black and white pixels at the same time.

3.3 First level classifier

Eleven invariants are used for feature extraction and classification process. Its sole job is classifying similar shaped characters into same sub group for later processing. Due to the scope of this report, only 50 characters need to be recognized, it is possible that first level classifier can produce final output directly for certain characters which do not require second level classifier.

3.4 Partial features finding

As features on whole character have already computed by first level classifier, second level classifier only needs to find important areas to further classify. Optimum grid is used to find interest areas and group specific second level classifiers are used to produce final output. Second level classifiers are a set of simple generic classifiers. For example: classifier for sub group “甲”, “田” and “由” is finding position of center of mass. Classifier for sub group “又” and “叉” is count number of group of connected pixels. Classifier for subgroup “己”, “巳” and “巳” is checking connectivity and so on.

4. Discussion and implementation

The whole OCR process can be considered as three main steps, preprocessing, processing, post processing. Each of the steps can be considered as a whole research area and has a lot of research potentials.

4.1 Pre-processing

Assume the original paper document has been successfully scanned into computer, preprocessing includes segmentation, binaries, noise reduction, normalize.

4.1.1 Segmentation

Segmentation is the foundation process in OCR process, without a good segmentation, the extracted character may useless for further processing and may occur unfixable errors. Segmentation can includes paragraph analyze, differentiate text and non-text contents and extract characters correctly from input.

For simplicity, we will consider the step extracts characters from input image as segmentation. A very simple segmentation method is projection, counting black pixels along X to differentiate different characters and along Y to split lines (figure 1.1).

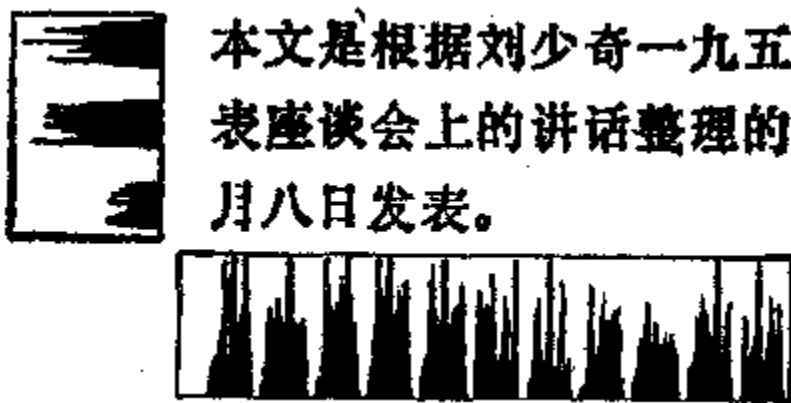


Figure 1.1 segmentation using projection adopted from X. Ding's Chinese Character Recognition[2]

However, problem can be raised if the paper was not placed on scanner properly (figure 1.2). How well is paper placed also determines error handling ability for classifier in later recognition process which will be discussed in later sections.

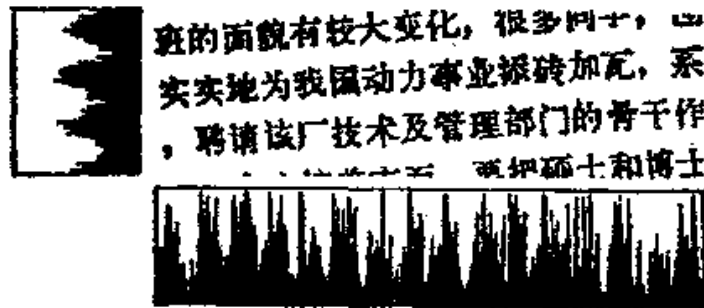


Figure 1.2 incorrect paper placement leads to segmentation error [2]

Problems can still occur even if paper has been placed correctly, consider figure 1.3 as input, the structure of Chinese character in big font can lead to false segmentation.

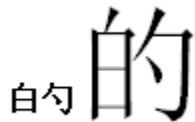


Figure 1.3 example input

As figure 1.3 shows, it is hard for computer to tell whether it is “白” and “勺” (left side small font) or single “的” (right hand side big font). So it is not safe to split character when there is low number of pixels represented on X axis, which is also mentioned in figure 1.2. Thus we could cut when there is an absolute gap between characters.



Figure 1.4 example input

However it is not the case. Figure 1.4 is from Fedora 10 Linux operating system using open office, when typing in Chinese characters, there is no absolute gap between characters and in this case, paper is perfectly placed. Even by cutting the first character manually, we get a dot on right side of the image (figure 1.5) which can significantly affect recognition process.

One solution is, combine manually selected font size and automatic validations after segmentation. Most input images will have similar size of fonts, thus if error rate for validating is not very high, it can assume it is a good segmentation.



Figure 1.5 dot left from next character

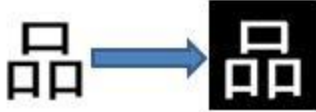
4.1.2 Binaries

In this report, to simplify the complicity and avoid any unnecessary pixels may affect the recognition process, simple threshold is used. This eliminates the smooth character affect provided by many operating systems including Windows and Linux and other kind of noises may have been introduced during input. Unlike human normally read black on white, binaries also need to convert characters to “white on black” as white is value 1 and black has value 0. This is an important step to cooperate with movement invariant classifier. As we are considering grayscale images as input, threshold used in this report is set at 128 and is defined as below for a pixel value at position (x,y) on image f :

$$f(x,y) < 128 \quad f(x,y) = 255$$

$$f(x,y) \geq 128 \quad f(x,y) = 0$$

Above equations will produce results shown in figure 3.1.2.1.



Original After binaries

Figure 3.1.2.1 binaries and reverse colour

4.1.3 Noise reduction

Binaries images usually still contain noise. One approach is applying median filter to achieve noise reduction. The advantage of median filter is it keeps the edges of the image but will eliminate some of the noise. Combine with binaries, problem in figure 1.5 no longer exists. (X.Ding, 2003) It is stated that apply Gabor filter is the best solution because it is robust and is very similar to human vision system.

4.1.4 Thinning

Strokes in Bold font and big font normally thicker than the others, it leads to a significant change in its features (figure 1.5.1 and 1.5.2), it worth noting that distance between 2 different characters in same font is smaller than distance between 2 different fonts for a same character. Thus a thinning is required for further processing. However thinning will change character features and researchers are trying to improve it. In this report, all fonts are not thinned and no bold font samples are used.

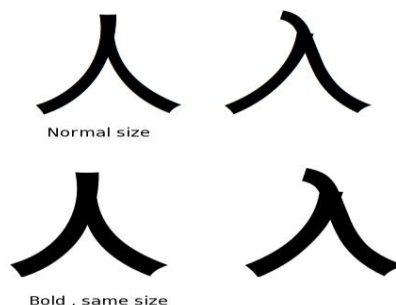


Figure 1.5.1 sample fonts

人	-6.5116	-22.6144	-44.6435	-29.3165	-47.6081	-32.5829	-12.4871	-23.5933	-37.0735	-58.3137	-60.3963
人	-6.3107	-25.0939	-47.0341	-33.4538	-48.0246	-33.9088	-12.0832	-22.5282	-37.1906	-64.1313	-64.8000
入	-6.2936	-23.2327	-45.0103	-32.3090	-44.3149	-31.1642	-12.0588	-23.4432	-38.0123	-60.2173	-59.8802
入	-6.5769	-22.3377	-44.2800	-29.4307	-44.6257	-29.7795	-12.5928	-21.7162	-35.7129	-58.2241	-58.2470

Figure 1.5.2 11 moment invariants from samples

4.1.5 Normalization

To achieve a good recognition result, normalize is used to scale and move the character to a normalized size and aspect ratio. Good normalize algorithms will benefit applying grid for further processing. According to X.Ding's research, normalize character according to its center of mass is more robust than according to its boundary because some of the part of character will be missing when input quality is not perfect [2](figure 1.6).

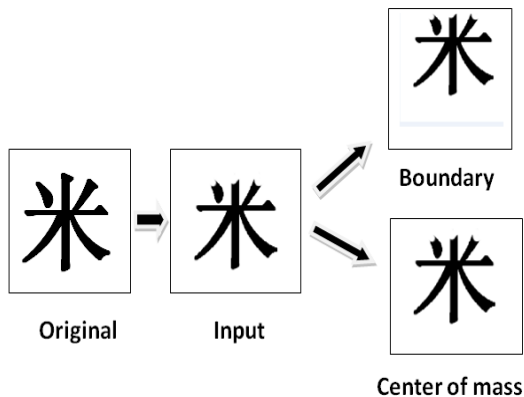


Figure 1.6 normalize according to boundary v.s. center of mass

To cope with different font size, normalization will also adjust font's size and aspect ratio to match kernel's size. Some fonts, normalization might change character's aspect ratio, thus character's feature might be changed and may lead to false detection. We are not considering such effect in this report.

4.2 Processing

Processing is the most important part of the whole process, it directly affect accuracy and speed of OCR procedure. Chinese character is the only graphic character used in ideography in the world and each of them has unique 2D shape representation. There are many algorithms can be used as described in literature review, and it is possible to combine them to achieve better results. This report will focus on the moment invariant method as a starting point.

4.2.1 The challenge

Assume the input and output from pre-processing are perfect, which means the fonts and size of the characters are expected, characters has been normalized correctly, noises are eliminated properly and there is no missing part for the character, we will only consider the there might be some small rotation transformations in input and the fonts are not thinned. The feature extraction step can still be challenging.

4.2.1.1 Hugh amount of characters and complex structure.

In GB2312-80[5], which is the most widely used character set in China, there are 3,755 for the most used characters. Its extension, GBK[7], covers 21,003 Chinese characters in Unicode. This means working recognition algorithms for English characters or other language scripts cannot be directly used on Chinese characters. Chinese character's structure varies a lot, history has proven that all algorithms that try to extract features based on Chinese character structure or strokes are failed. Similarly, any algorithm generates high dimensions will not be ideal for Chinese characters as computational time will be high, it will make training process extremely slow as well.

4.2.1.2 Similar characters.

In numerical recognition, there are only two similar shaped numbers, "6" and "9", in English characters, there is none. However, there are more than 200 of the characters are very similar shaped characters [2] with totally different meanings

and pronunciations. For example “人”and “入”, “又” and “叉”.Some of them even are not differentiated by one completely stroke “巳”, “巳” and “己”. This made feature extraction very challenging. As described before, any classifier algorithm perform based on whole characters will either eliminates or shorten distance between different similar geometry shaped characters thus most of one level classifier method do not suit Chinese character recognition.

4.2.2 Moment invariants classifier

To deal with such huge amount size of character set and in order to get a big enough difference between each characters with reasonably speed, a good and robust algorithm to extract features is desired. The nature of Chinese character is each of them has its own unique geometric shape, thus moment invariant can be applied to the feature extraction step, the most important step in OCR process.

4.2.2.1 The seven HU moment

The 7 HU moments, was originally proposed by Hu in 1962. The moment invariant values of the image/character are independent to position, scaling and rotation. It can be used in geometrical patterns and alphabetical characters reorganization [1]. Many text books for image processing presented the equations for clarity such as Gonzalez and Woods. OpenCV also have this method build-in.

The moment order (p+q) for a 2D image is defined as:

$$m_{pq} = \iint_{-\infty}^{\infty} x^p y^q f(x,y) dx dy$$

The central moments for digital image are defined as:

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q f(x,y) dx dy$$

Where

$$\bar{x} = \frac{m_{10}}{m_{00}}$$

and

$$\bar{y} = \frac{m_{01}}{m_{00}}$$

The normalized central moments η_{pq} are defined as

$$\eta_{pq} = \frac{\mu_{pq}^Y}{\mu_{00}^Y}$$

where

$$\gamma = \frac{p + q + 2}{2}$$

In figure 4.0.1, data is calculated based on same font and size on both character but rotate them in varies ways, it clearly shows it is possible to extract features using 7-HU method and this algorithm is very robust to rotations even there are some rounding errors when doing transformations.

However in Chinese OCR, the rotation invariant feature of moment invariants can affect accuracy because there are characters “rotational sensitive” like “甲” and “由”.

Furthermore, in this classic 7-HU method, the 7 features are not truly independent as they should be. Researchers also want more features to be extracted using this algorithm to enhance accuracy and get bigger distances.

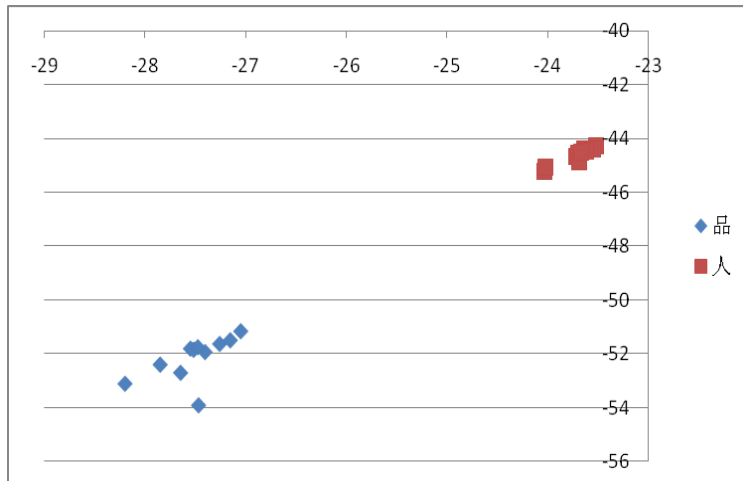


Figure 4.0.1 HU moments, Φ_4 vs Φ_5

4.2.2.2 Flusser third and fourth order of moment invariants

Although there were many attempts had been done related to expand Hu’s system, Flusser[3] proved that none of expanded algorithm is truly independent

invariants in 2000. Furthermore, his work proposed that the 3rd order in HU's moment invariant is dependent and there are two third order invariants are not included in Hu's system. To cope with similar shaped Chinese characters, one possible solution is using higher order moments in order to extract more features in the hope of getting big enough distances between these characters. However, Flusser suggests value for p and q should be as small as possible because in Teh and Chin's work[9], they state higher order moment invariants are more sensitive to noise than lower order ones which will generate false features. Furthermore, according to curse of dimensionality, training and computation time will slow the OCR process down by considerable amount thus order above fourth order is not recommended. Barczak[4] extends the moment invariants further up to fourth order in working equations. Flusser proposed that there are eleven independent elements up to the fourth order. Although higher orders amplify the difference, they are still not ideal for Chinese OCR. Figure 4.1.11 shows Ψ_7 and Ψ_8 of moment invariants for “人” and “入” are completely overlapping. As expected, problem of small distances between similar shaped characters remains (in fact, in this certain example, these two characters have some big distances in other orders. However, as number of samples increase, the distance between them becomes very small. Furthermore, in order to make algorithm on whole Chinese character set, not limited by 50 characters in this report, following discussion is needed, in case there might be some characters cannot be recognized by movement invariant solely.)

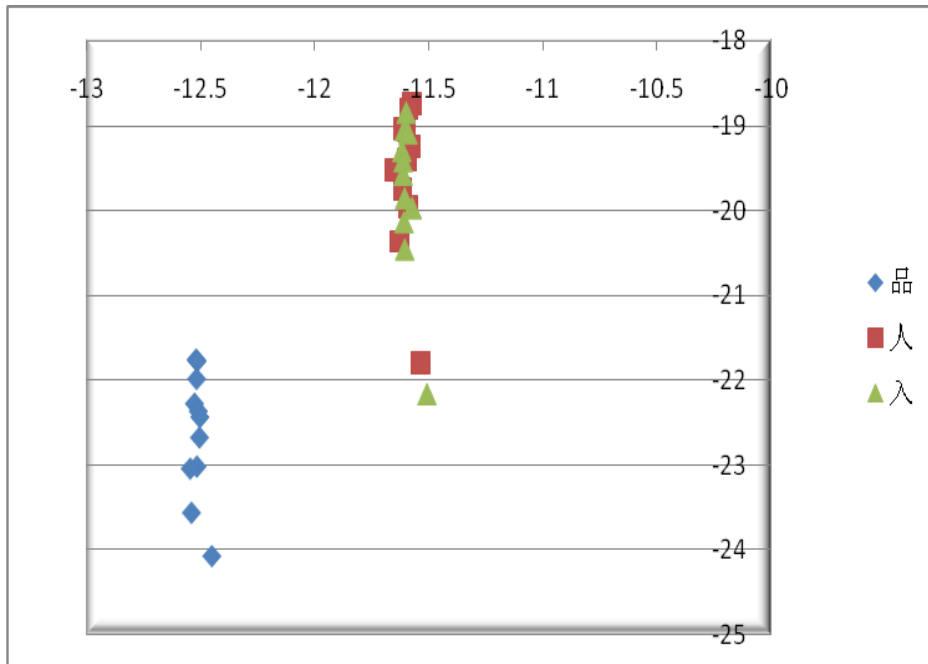


Figure 4.1 eleven moment invariants from characters ψ_7 v.s. ψ_8

4.2.3 Two level classifier architecture

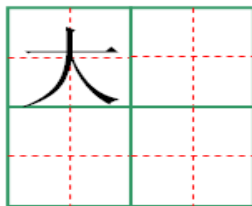


Figure 6.1- “田” shape grid

In literature review, it is recommended that split the character recognition step into two: classify similar characters into one group using first level classifier then apply further feature extractions to get desired output using second level classifier (X.Zhu et al 1987).

The rule of thumb for first level classifier (sorter) is it should be robust enough to cope with small rotations; noise and missing strokes for pre-processed characters and it should be efficient that does not slow down the whole OCR process. A ideal performance for first level classifier should be: the distances between each subgroups should be as big as possible and the distances between members in

one subgroup should be as small as possible (X.Ding, 1992). In this case, moment invariant makes a perfect method to match requirements suggest by Ding as first level classifier. We can make use of the short comings in movement invariants algorithm as first level classifier and it can sort similar shaped characters in same sub groups, for example: group one contains “人”, “入” and”八”, group two contains “己”, “巳” and “巳” group three contains “甲”, “田” and “由” and so on. Particularly, for subgroup “又” and “叉”, it also provides robust to noises, i.e. the middle dot in “叉” might be noise.

To find how to find more detailed features for second level classifier, it is good to know how human beings learn writing Chinese, they write characters in a “田” (pronounced tian) shape grid (figure 6.1). One approach is applying grid to extract detailed features from the character. Umeda[7] and Akiyama[8] mentioned is divide the image into parts according grid and then calculate detailed features from each part of the grid like density of black pixels, distance to edge of the grid from a chosen points using different methods. This raises complicity of the problem and not efficient because of the huge amount of dimensions. (L.Jin et al,1997) Using dynamic grid to dynamically decide density of the grid that how the character to fit in, i.e. areas with high density of strokes should apply a higher density of grid and verse-visa. Despite what grid method is used, the main disadvantages are grid method is very sensitive to 2D transforms; a small shift in the position can cause significant errors in calculation. And grid method often very computational expensive and generates lots of dimensions. High dimension problem will lead to time consuming training as discussed before. As features have already been calculated on whole character, it is obvious that no need to calculate features based on whole character again. By studying human vision system and human reading’s procedure, only a small area needs to be computed in each subgroup in order to get final output. The solution in this

report is using an optimum grid to calculate detailed features for interested areas only. The advantage of this idea is it is big enough to keep robust nature (if there is any) of the algorithm for further processing and small enough to save computation time and do computations only when it is needed to. For example, we choose the top half of the character for “入” and “人” as interested areas, marked as red areas in figure 6.2. Of course in this example characters are normalized in center and blank margins appeared which is not as same as practical problems. But the idea of optimum grid stands.

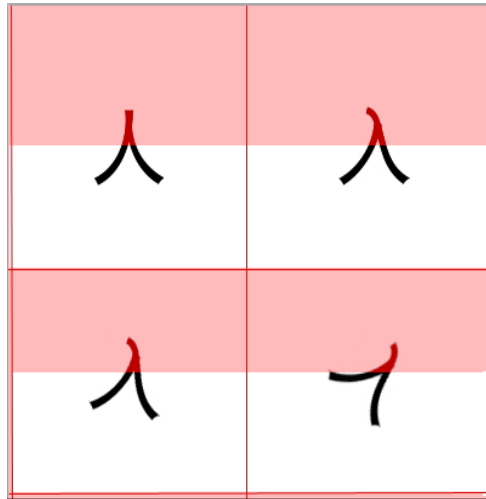


Figure 6.2 choose interested areas

Therefore the algorithm becomes as shown in figure 6.3. Moment invariants method is used as a first level classifier that groups similar shaped characters for further processing.

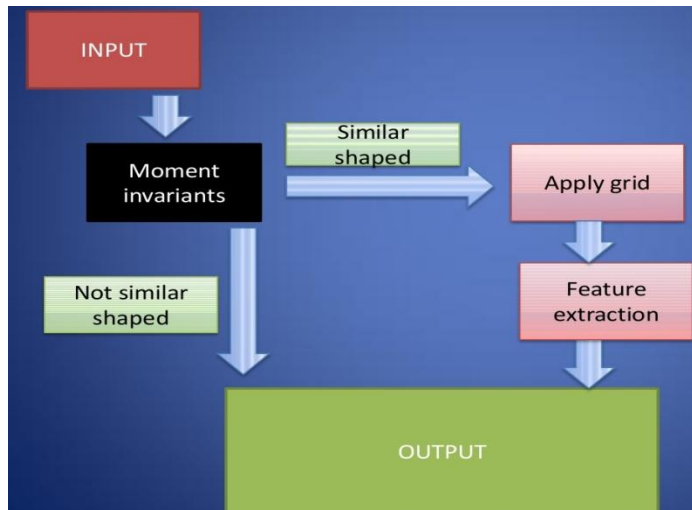


Figure 6.3 OCR processing algorithm

4.2.3.1 Apply Optimum Grid

The selection of interested areas should be same for same group of characters. In implementation of finding interested areas, there are two possibilities: dynamic locate and hard code. Due to complicity of structures in Chinese characters, there is no known efficient way to find interested sub windows by algorithms. There are about 220 individual characters in Chinese needs to be grouped. Normalizing procedure should adjust character to fill processing kernel, hard code boundaries according to kernel size can be a solution.

4.2.3.2 Further feature extraction

Find a generic efficient second level classifier is challenging. As the size of each character is 96*96 and optimum grid will reduce this size considerably (because they are grouped with similar shape and only a small amount of areas can tell the difference), more computational expensive calculations can be applied without affecting whole program efficiency. At least second level classifier should have comparative advantages compared to first level classifier. In this report, heuristics of similar shaped character such as center of mass, connectivity are used.

Figure 4.2.3.2.1 demonstrates detailed work flow. Input characters with similar shapes are classified by movement invariants into same subgroups; optimum grid has been applied, using 1/3 of total image size in this case. Center of mass as second level classifier is then used to further classify. Center of mass has some robustness to rotations and this feature has been kept after applying optimum grid.

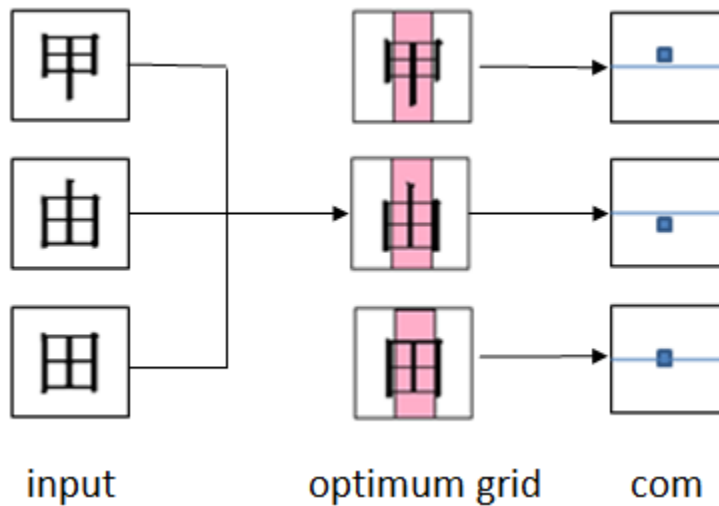


Figure 4.2.3.2.1 simulation of optimum grid with second level classifiers

4.3 Neural networks

Neural network is another simulation on human vision which plays the role of human brain. It takes a lot of samples and gets a meaning of each sample and once it is trained, it can take characters in unknown size/font and recognize them. Feed forward back propagation neural networks are widely used in character recognition area. However design a neural network can only trail on error. There is no way to determine number of neither hidden layers nor hidden nodes in the neural network for a specific problem. Fortunately, it is proven (R.Hechi-Nilson,1989)that three layer back propagation neural network are suitable for this problem. The neural network in this report uses 50 output nodes, 11 input

nodes and 50 hidden nodes. All input nodes are connected to hidden layer nodes, all hidden layer nodes are connected to output nodes as shown in figure 3.3.1. Apart input nodes, all other nodes have one bias node with output always equals to 1.

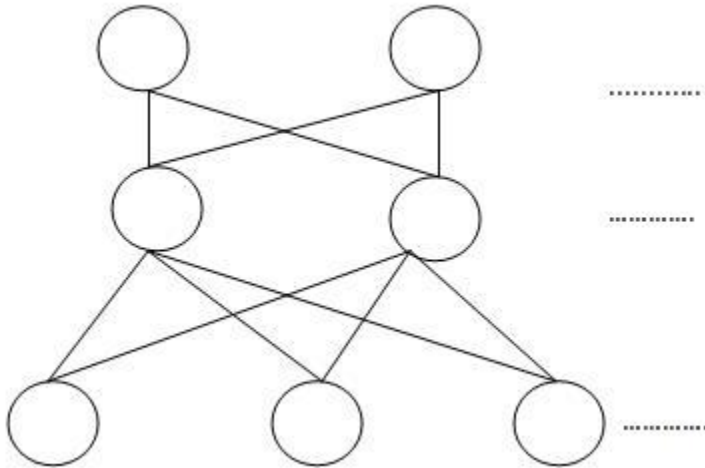


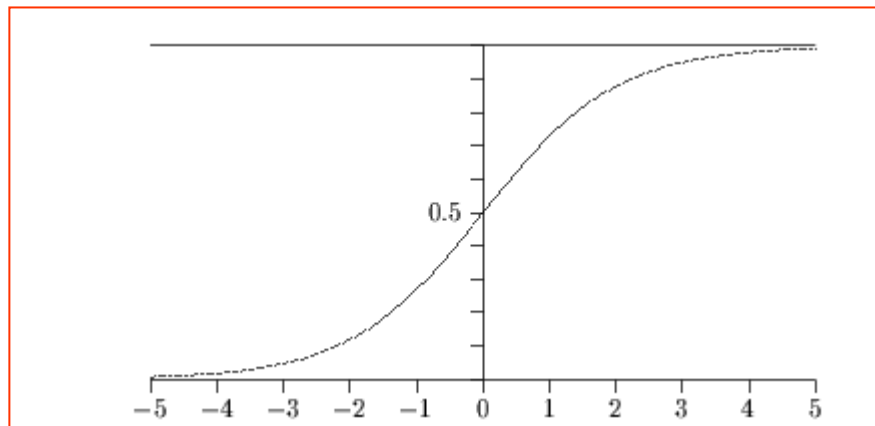
Figure 3.3.1 neural network setup

As character recognition is non-linear problem, we use the most commonly used non-linear function

$$v = \frac{1}{1 + e^{-s}}$$

standard sigmoid:

this function looks like the following chart:



the activation value for a neuron j is: $O_j = f(net_j)$

the net input to neuron j is: $net_j = \sum_{i=1,n} W_{ij} O_i$

First of all, we need to train the neural network, as it is supervised training, each input and output are in pairs, the error signal for output node k is:

$$\delta_k = (t_k - o_k) f'(net_k)$$

Where if we use usual activation function $\frac{1}{1 + e^{-net_k}}$

The derivative term is $o_k(1 - o_k)$

Thus the error signal is now:

$$\delta_K = (t_k - o_k) o_k (1 - o_k)$$

Where t_k is target output of k and o_k is actual output.

All weights between neural are randomly initialized to (0,0.1), in training process, weight between two neural j and k is adjusted using:

$$W_{jk} \leftarrow W_{jk} + \eta \delta_k o_j$$

Where η is learning rate set to 0.1 in this report.

Training process was as N.Reyes[24] suggested in his lecture notes as follows:

1. Three sets of characters are spitted. Because each of the character has 5 different fonts, 3 of them are used as training set with total number of 1,500 samples, 1 of the font is used as validation set to stop training and the last font is used as test set to evaluate performance.
2. All weights are randomly initialized from 0 to 0.1 non inclusive.
3. Training stopped when error signal is smaller than 0.1.
4. Evaluate performance.

These also followed his suggestion that training set should include rotation and variation in font style.

4.4 Post processing

Not all characters can be recognized or recognized correctly by program. As we need to consider computational time, system resources and accuracy, there are always a blind spot for one certain character recognize algorithm (X.Ding et al,1995). Corrections can be made according to the meaning of the sentence, word and the frequent use of the character. Post processing can enhance accuracy. For example “人们” means people while “入们”does not mean anything, post processing can fix such errors which would be very difficult for algorithm to do otherwise. The nature of Chinese characters actually makes recognition much easier than other language (X.Ding, 2003). However, in this report, no post processing is used.

5. Results and analysis

Two sets of results are obtained in this report. During the progress of this project, recognition purely based on movement invariant was developed first. Then optimum grid with second level classifier has been tested.

One Classification based on movement invariants:

95.2% of the characters are recognized correctly.

Not surprisingly that 甲由, 日日 and 己己巳 have misclassification rate of 20-35% in corresponding subgroups. While using two level classification methods we achieved 99.4% accuracy. Time increased for each character which requires second level classification is less than 1ms. Robust second level classifiers recognized all characters correctly which otherwise would be misclassified, I believe with an efficient classifier, performance can be further improved.

It worth noting that movement invariants classified all 50 characters into correct corresponding subgroups which have achieved 100% accuracy. However, it can take up to 5 ms to finish movement invariant values.

To deal with multi-font printed Chinese character recognition, the following characters in figure 5.1 generates the most errors in one level classification method. After apply “日” font Hei and “日” font MS Yahei, these two characters are trend to be very similar and it is hard to recognize even by human beings.

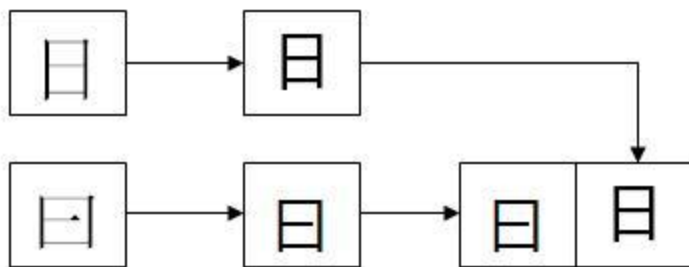


Figure 5.1 different characters become very similar in different fonts

However it is possible to recognize by algorithm introduced in this report. By checking number of groups of connected pixels on right half of each character, 日 will give output 1 and 日 will output 2. A detailed table in figure 5.1.2 showed as below:

Type of second level classifiers	Example subgroups	Robustness	Example of failure
Center of mass	甲田由	robust	Failed at big rotated angles
Connectivity	己 己 巳 (purely based on connectivity)	Not robust	Failed at any rotated angles and some fonts
Number of groups of connected pixels	己 己 巳 (with connectivity), 日 日	Very robust	Very few failure

Figure 5.2 type, robustness and application for second level classifiers

Checking connectivity has the worst robustness on rotated characters mainly because of false determination of horizontal strokes (I used simple checking on pixels on X position), however this might be enhanced by line detection. This again proved that trying to analyze strokes of characters is not ideal for Chinese OCR.

Center of mass is not very robust to big rotated angles, it is might not be robust to transformations neither depending on the structure of the character. However, in practice, characters should not be as much as been rotated by 10 degrees so center of mass method can still provides some robustness if normalization has been done correctly.

Combination of connectivity checking and number of groups of connected pixels are used to check subgroup ㄱㄱㄱ, experimental result shows it can recognizes all “ㄱ” but it increases misclassification rate between “ㄱ”and “ㄱ” if features from movement invariants are discarded. In another word, to achieve better accuracy without affecting speed, features from first level classifier should be reused.

Failure caused in number of groups of pixels are by connectivity checking between ㄱ and ㄱ, otherwise there is no failure caused in that classifier and we can conclude that optimum grid keeps robustness nature of classifiers.

Of course accuracy using two level classifications can be improved further by enhancing second level classifier, as classification accuracy of movement invariants is 100%. Due to the scope of this project, only 50 characters in 5 fonts are to be recognized and they are all perfectly pre-processed, methods used in this report is relatively efficient.

6. Conclusions

6.1 Summary

Experimental results show that optimum grid can keep robust nature of classifier and will reduce size of image to be processed by a considerable amount. In a multilayer classification OCR architecture, features from previous layers should be reused to achieve better accuracy without slowing performance. Furthermore, classifiers should have complementally advantages as well as some robustness to cope with transformations and noise and in practice.

Commercial Chinese character recognition developed by Tsinghua University is now robust and high performance, it recognizes over 20,000 Chinese characters as well as over 6,000 Japanese characters and over 7,000 Korean characters. These also cover mixture of above languages and English characters. Now it can handle near 100 various fonts and low quality printings.

It also solved samples have huge different transformations problem. Online hand writing recognition rate is 98% and offline hand writing recognition at 92%. They used statistical reorganization which makes Chinese characters easier to analyze than other western characters which methods based on structure analyze will not be close in performance. Complex layout of input image analyze are also solved. Now it is wildly used in China and saves huge amount of labors.

However, OCR is still an open area for computer science researches especially in low printing quality inputs, big transformations, complex noises, to recognize characters in video sequence, character's font, signature recognition.

By using optimum grid, potential super linear speed up can be achieved compare to normal grid method while optimum grid keeps robust nature of feature extracting algorithm.

6.2. Limitations

Every step in this report can be extended to a whole field of research. Due to the scope of this report, it mainly focuses on processing step. We assumed pre-processing is perfect, but in practical it is impossible. There is always going to have characters with noises or missing part, all commercial OCR products has relatively low performance if input image is in low quality compared to performance on high quality input. Character font and size are not always as expected in practice, this will cause false segmentation and leads to un-fixable errors as discussed. As we are dealing with less than 0.25% characters in Chinese character set, performance on uncovered characters are unknown. These all can decrease accuracy of the program in practice and we expected as a not-small amount. Second level classifiers in this report are also subgroup specified, it is not safe to predict whether this method will work in practical; however we can always create new second level classifiers for new subgroups. Although we did training part of the system, we are only dealing with 15 characters in 5 fonts with same size. Our training sample is small and training time is limited, this can directly affect accuracy as well. A large enough training sample, which covers all possible fonts and covers hopefully all sizes, will increase accuracy significantly, especially for printed character recognition, eventually it will become as a acknowledge based system as test set is as same as training set.

6.3. Feature work

This is a potentially big project, with limitations discussed above, in the future I would like to extend number of characters to be practical and perform extended test on optimum grid algorithm.

7.References

- [1] HU, M.K. "Visual pattern recognition by moment invariants" , *Information Theory, IRE Transactions on Information Theory*, Vol.8, Issue 2, pp.179-187, Feb. 1962.
- [2] Ding,X. "Chinese Character Recognition: Principles and Implementation Methods",1992.
X.Ding "Chinese Character Recognition: A Review", *Acta Electronica Sinica*, Vol.30, No.9, Sep. 2002
- [3] Flusser, J. "On the independence of rotation moment invariants", *Pattern Recognition*, Vol.33, pp.1405-1410. 2000.
- [4] Huang,Y.&Suen, C. " A method of combining multiple experts for the recognition of unconstrained handwritten numerals", *IEEE Trans on PAM I*,pp90-94, 1995
- [5] Suen,C.&Legault,R."Building a new generation of handwriting recognition system". *Pattern recognition Lett*,pp.303-315, 1993
- [6] Ho, T.,Hull, J.&S.Srihari,"A computational model for recognition of multfont word images",*Machine Vision and Application*, pp.157-168,1992
- [7]China National Information Technology Standardization Technical Committee, "GBK",1995.
- [8] Barczak, A. "Feature-based Rapid Object Detection: From Feature Extraction to Parallelisation", *a thesis presented in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Computer Sciences at Massey University, Auckland, New Zealand*.pp.105-109.2007.
- [9] Amheim,R."Visual Thinking" .*University of California press*,1969.
- [10]L.Jin & Z.Xu "A new Feature extraction method: elastic meshes features", *Journal of Circuits, system, and computers*, pp7-12, vol.32, 1997
- [11]R.Hechi-Neison,"Theory of the back propagation neural network" *Neural networks,1989.IJCNN.,International Joint Conference on*, 1989
- [12]Papageorgiou, Oren and Poggio, "A general framework for object detection", *International Conference on Computer Vision*, 1998
- [13] Viola, P. & Jones, M., "Rapid object detection using boosted cascade of simple features", *Computer Vision and Pattern Recognition*, 2001
- [14]Messom,C H. and A.Barczak,"Fast and Efficient Rotated Harr-like Features Using Rotated Integral Images", *Australian Conference on Robotics and Automation(ACRA2006)*, pp.1-6,2006
- [15]State Bureau of standardization of the People 's Republic of China, "GB2312-80" published in 1980 and became effective on 1 May,1981.

- [16]X.Zhu,Y.Wu&X.Ding,“The recognition of 6763 Printed Chinese Characters”,*Journal of Tsinghua University*, Vol.27,No.1,1987
- [17]M. Umeda “Classification of Multi-Font Chinese Characters”, Trans. Of IECE of Japan PRL-78-53,1978.
- [18]X.Zhu, Y.Wu and X.Ding, “A New method for extracting radicals of printed Chinese characters”*Journal of Computer Science*, Vol. 11, 1988.
- [19]T.Akiyama&N.Hagita “Automatic Entry System for Printed Documents”, *Pattern Recognition*, Vol.23, No.11,pp.1141-1151,1990.
- [20]C.H. Teh, R.T. Chin, “On image analysis by the method of moments”, *IEEE Trans. Pattern Anal. Mach. Intell.* 10,pp.496-513,1988
- [21]X.Ding&F.Guo,“Chinese character recognition: achievements”,*Software Inside*, 1995
- [22]Flusser,J.&Suk, T. “Pattern recognition by affine moment invariants”, *Pattern recognition*, Vol.26,No.1, pp.167-174,Elsevier Science,1993.
- [23] Y.Liu “Printed Chinese character recognition”, *a thesis presented in partial fulfillment of the requirements for the degree of Honors of Science in Computer Sciences at Massey University, Auckland, New Zealand*, 2009.
- [24] N.Reyes,“Neural networks”, *Studies in Machine Learning*,2009.

Appendix

Raw moment invariants for characters used in figures and tables

All characters are in same font and same size, picture size is 200x200 pixels
 Last column is in format of **Character number** where number is the rotation degree.

For example: 品 -10 means character 品 rotated by -10 degrees

-6.41537	-27.26210	-51.62660	-37.20330	-51.82010	-35.38990	-12.52160	-21.74940	-37.53620	-69.05420	-68.67440	品-10
-6.41814	-27.65060	-52.69950	-38.33100	-52.08290	-35.74600	-12.52920	-22.27040	-38.18900	-71.92820	-69.29580	品-5
-6.40755	-27.55330	-51.79960	-37.03410	-57.55910	-37.24240	-12.50750	-22.67130	-38.46780	-69.06820	-75.17190	品-3
-6.38002	-27.40730	-51.92810	-37.02750	-51.84610	-35.61730	-12.45360	-24.07500	-40.15780	-68.63960	-72.50910	品
-6.41451	-27.47170	-53.90860	-39.44350	-51.68480	-35.92500	-12.51850	-23.01480	-39.05650	-69.69050	-68.99030	品 3
-6.41106	-27.85420	-52.39470	-37.77870	-52.92920	-37.44040	-12.51270	-22.36210	-38.48500	-69.92130	-70.08210	品 5
-6.41403	-27.52050	-51.83970	-37.31050	-52.63820	-38.10800	-12.51770	-21.76730	-38.02320	-69.18730	-69.54990	品 10
-6.41589	-27.15620	-51.49040	-37.21930	-51.60710	-35.43730	-12.52030	-21.97860	-37.74470	-68.87170	-68.44000	品 7
-6.41075	-28.20370	-53.10610	-39.08640	-53.01610	-37.33990	-12.50580	-22.42960	-38.72200	-71.10760	-70.55470	品-7
-6.42501	-27.47920	-51.74540	-37.13020	-52.91410	-37.32970	-12.54240	-23.56170	-39.57600	-69.00420	-69.72670	品 2
-6.42544	-27.05230	-51.15340	-36.81920	-52.28760	-36.44650	-12.54750	-23.03960	-38.94170	-68.18530	-68.94010	品-2
-6.04937	-23.64050	-44.39990	-31.59750	-46.95480	-31.65920	-11.57220	-18.73120	-32.83980	-60.41820	-62.13320	人-10
-6.05015	-23.51380	-44.28430	-31.46490	-45.20430	-34.33910	-11.57500	-19.23790	-33.31540	-60.24060	-60.80500	人-5
-6.07194	-23.59380	-44.41000	-31.53730	-45.55760	-33.34680	-11.61240	-19.74410	-33.84170	-60.41000	-61.08180	人-3
-6.02758	-23.71570	-44.66350	-31.66020	-45.06200	-31.81030	-11.53270	-21.80250	-35.76310	-60.89250	-60.85930	人
-6.05955	-23.70000	-44.54640	-31.59360	-45.75390	-32.25790	-11.58850	-19.95320	-34.02090	-60.59160	-61.39910	人 3
-6.05935	-23.54710	-44.39810	-31.55430	-45.02370	-31.92210	-11.59310	-19.39330	-33.57560	-60.43750	-60.72170	人 5
-6.05584	-23.61710	-44.49420	-31.58980	-45.14470	-32.28270	-11.58550	-18.79380	-32.97220	-60.55320	-60.88890	人 10
-6.06142	-23.68370	-44.87060	-31.76770	-44.78690	-31.78250	-11.59590	-19.06110	-33.23760	-61.22190	-60.66400	人 7
-6.07068	-23.66520	-44.51520	-31.70400	-45.60940	-33.29420	-11.61040	-19.03040	-33.18300	-60.60920	-61.29600	人-7
-6.07687	-24.03040	-45.22550	-32.03380	-45.58230	-31.99210	-11.62780	-20.36160	-34.43300	-61.63340	-61.62060	人 2
-6.08658	-24.01960	-45.06120	-32.10130	-46.28870	-34.71070	-11.64960	-19.51560	-33.69690	-61.43460	-62.10920	人-2
-6.06919	-22.25310	-43.35680	-30.81900	-42.47130	-30.14260	-11.60580	-19.04750	-33.14180	-57.81220	-57.33140	入-10
-6.07586	-22.26280	-43.47760	-30.81500	-42.47920	-30.09730	-11.61620	-19.29550	-33.38740	-57.93000	-57.33720	入-7
-6.07202	-22.17640	-43.08090	-30.80540	-42.40020	-29.92060	-11.61150	-19.57670	-33.63450	-57.51510	-57.23000	入-5
-6.05509	-22.20860	-43.02150	-30.87730	-42.43930	-29.85860	-11.57290	-19.97090	-33.98080	-57.49360	-57.31430	入-3
-6.06683	-22.22270	-43.15200	-30.78650	-42.45750	-29.95040	-11.60420	-20.45950	-34.51270	-57.61150	-57.34460	入-2
-6.02089	-21.99630	-42.68400	-30.50860	-42.05550	-29.80950	-11.50930	-22.17050	-36.45220	-57.01350	-56.81900	入
-6.06811	-22.12330	-43.08220	-30.68060	-42.29380	-30.10160	-11.60780	-20.13910	-34.32180	-57.45830	-57.07990	入 2
-6.06689	-22.26530	-43.00400	-30.99480	-42.58740	-30.16460	-11.60490	-19.85600	-34.01110	-57.56510	-57.56760	入 3
-6.07231	-22.15940	-43.09580	-30.77370	-42.36330	-30.32920	-11.61150	-19.41700	-33.61390	-57.51490	-57.15550	入 5
-6.06095	-22.23190	-43.26050	-30.89080	-42.43890	-30.55320	-11.59060	-19.08660	-33.24380	-57.75030	-57.26440	入 7
-6.06463	-22.16340	-43.07360	-30.74920	-42.36030	-31.13320	-11.59850	-18.84090	-33.11470	-57.49560	-57.19040	入 10

