毕业设计（论文）题目：图像中的文字识别算法研究

学　　院：信息与电子学院

专　　业：电子科学与技术（全英文教学专业）

班　　级：05931201

姓　　名：高　暄

指导教师：李慧琦

摘　要

图像信息可谓无处不在，我们通过视觉读取外界信息。随着科技的进步，人们越来越需要利用图像上的信息，尤其是文字信息来识别和判断事物，并解决实际问题。同时更重要的是对文字图像进行处理，对图像处理技术也提出了更高的要求。文字识别也不只停留在英语字母和数字上，对汉字识别的要求也日益提高。

本文的文字识别是指计算机自动识别文字，它的主要技术是 OCR （Optical Character Recognition，光学字符识别）。本毕业设计项目的目的就是利用 OCR 文字识别算法识别特殊用途的有背景的图片中的文字信息，如龙标和教辅用书封皮上的信息。本文在阅读大量相关文献和充分调研的基础上，阐述了 OCR 文字识别算法的原理，并用 MATLAB 语言完成对特定图片文字信息的识别。

在论文安排上，首先，介绍了论文背景和调研结果；其次，较深入地阐述了 OCR 文字原理；然后，分别详细讲述了在 MATLAB 上如何实现对龙标和教辅用书封皮图片上关键文字的识别；接着对实验结果进行了分析和主观评估。

关键词：MATLAB；文字识别；图像处理；OCR

Abstract

Text in images is the main source for human to obtain the information from the outside world. At present, more and more people using image information to judge things and solve practical problems. Therefore, the text information in images is very important. Meanwhile, it is more important to deal with the text image, and put forward higher request to the processing technology of the character recognition. Text information is not only in English letters and numbers, but also Chinese character recognition.

Text recognition uses the computer to automatically identify the text message. Its main technology is OCR (Optical Character Recognition). The purpose of this graduation project is to use the OCR algorithm to identify the text information in some special pictures, like “Longbiao” and reference book cover. On the basis of reading a large amount of relevant literature and investigation, this paper expounds the principle of the OCR text recognition algorithm, and uses MATLAB to complete the recognition of the key information.

Firstly, this paper introduces the background and research results of character recognition; secondly, deeply expounds the principle of OCR; then, a detailed account of how to realize the character recognition implementation of special requirements for the images in MATLAB. Finally, the experimental results are analyzed and subjective evaluated.

Keywords: MATLAB；character recognize；image processing；OCR

Content

Chapter 1 Introduction

1.1 Background

Text in images is the main source for human to obtain the information from the outside world. In modern scientific researches, military technology, industry and trade of agricultural production, medicine, astronomy and meteorology, and other areas, more and more people using image information to judge things and solve practical problems. So recognizing text from image content is very important. It is more important to process the images and text to find the information we need. So in the era of today's rapid development of science and technology, higher requirements are needed in technology of text recognition. We need to get useful information more rapidly and higher accurate.

Text recognition is an important part of a new generation of intelligent computing interface. It involves the computer digital image processing, pattern recognition, artificial intelligence, fuzzy mathematics, combinatorial mathematics, information theory, natural language understanding and other disciplines. Optical Character Recognition (OCR) refers to use computer to automatically identify printed and handwritten words on paper or other medium. This graduation project is to study the character recognition algorithm, and help advertising company and tutoring company to do other researches. [1]

There are also more and more requirements about character recognition. For example, the bank needs to record all the data about depositors; the advertising company needs to know how many times people watch their ads. At the same time, the character recognition developed so fast, that it becomes to the deep learning period.

1.2 Application Area

Digital image processing is mainly to modify or improve the quality of the image, or extract

information from the image, as well as the use of digital image processing can reduce the size of images for transmission and preservation. At present, digital image processing is mainly used in communications technology, space exploration of remote sensing technology and bio engineering and other fields. Digital image processing is easy to achieve non-linear processing, and processing procedure and processing parameter are variable. It also has characters such as strong versatility, high precision, flexible processing, reliable information storage, transmission and image processing. It mainly used for image transform, measurement, pattern recognition, image generation and simulation. It also widely used in remote sensing, space observation, medical imaging and various industrial fields.

Character recognition can also be used in many fields, such as reading, translation, literature retrieval; data entry for business documents, e.g. check, passport, invoice, bank statement and receipt; automatic number plate recognition; extracting business card information into a contact list; more quickly make textual versions of printed documents, e.g. book scanning for Project Gutenberg; make electronic images of printed documents searchable, e.g. Google Books; converting handwriting in real time to control a computer (pen computing); defeating CAPTCHA anti-bot systems, though these are specifically designed to prevent OCR; Assistive technology for blind and visually impaired users; letters and parcels sorting; manuscript editing and proofreading; invoice summarizing, commodity code identification, goods warehouse management, and water, electricity, gas, rent, insurance and other personal expenses automatic processing and typist office work automation, etc.; all kinds of documents retrieval and certificate recognition. It improves the efficiency of all walks of our daily life. [2]

1.3 State of the Art

Text processing was first used in 1920s. In mid 1960s, the computers are widely used. Text and image processing technology has been improved, gradually it become a new science. From the beginning of the mid 70's, with the rapid development of computer technology, artificial intelligence and cognitive science, digital image processing technology go to a

higher and deeper level. In 90s of 20 century, people have begun to study how to use the computer system to understand images, which is similar to the human visual system to understand the outside world. That is known as computer vision. Many countries, especially developed countries invest more manpower and material strength for this study, and many important results are obtained. [3]

Early optical character recognition can be traced to telegraphy and creating reading devices for the blind. In 1914, Emanuel Goldberg developed a machine that can read characters and converted them into uniform telegraph code. At the same time, Edmund developed the Optophone, which is a blind reader and a handheld scanner that when moved across a printed page, produced tones that recognize to specific letters or characters.

In the early 60's and 70's, the study with OCR began all over the world. And in the early time of the study, the mainly researches are about text recognition method, and only recognize numbers from 0 to 9. For example in Japan, where also has square characters like China, began to study the basic theory of OCR recognition from about 1960. Initially, the objects are numbers. Then until 1965 to 1970, Japan arose some simple product, such as printed zip code recognition system. It can identify the zip code to help post office for regional distribution.

In 1978, Kurzweil Computer Products began to sell a commercial version of the optical character recognition computer program. In the 1970s, China began to study about numbers, letters and symbols recognition. At the end of the 1970s China began to study on Chinese character recognition. In 1986, China put forward the "863" advanced technology research program. The research on Chinese character recognition goes into a substantive stage. Prof. Xiaoqing Ding from Tsinghua University and the Chinese Academy of Sciences respectively did the research, and launched a Chinese OCR Products, which is now the advanced Chinese character OCR technology. After entering the 90's of the 20th century, along with the widely use of type scanner, and popularize of information automation and office automation, the development of OCR technology is greatly promoted. OCR

recognition correct rate and speed can meet the requirements of the majority of users. Various commercial and open source OCR systems are available for most common writing systems.

1.4 Confronting Problems

Now for the character recognition technology, it is also facing some problems. One is a large amount of image data. The second is image fouling. Fouling image can be due to the interference of environmental objectives, transmission error, sensor error, noise, background interference and distortion. The third is accuracy. For example, displacement, rotation and scale change, distorted. Such as human vision, there will be changes between the target and the sensor. The fourth is we need to handle on time. In military application, it requires a system capable of real-time identification of the target, which requires the system has fast speed and high efficiency of identification.

1.5 Missions and Meanings

1.5.1 Missions

Text in an image can provide important information for image understanding. Automatic identification of text information can facilitate image retrieval and image understanding. Great progress has been achieved in the field of OCR. But there are lots of challenges in identification text information from image for practical applications. For example, the background of the text is more complicated. And sometimes the information of color, intensity and contrast is not consistent. The objective of this graduation project is to investigate algorithm to identify text information in an image. According to this graduation project, one can strengthen knowledge of image processing and pattern recognition. In this graduation project, I need to study on two specific types of images, and get the key information by character recognition. There are “Longbiao” images and “reference book cover” images. I will analysis each type of images separately in detail later.

The missions of this graduation project are as follows:

(1) Literature review: Study the methods of text recognition. Write a report of literature review with citations of more than 20 research papers and describe the state of the art.
(2) Investigate the methods of text recognition. Propose algorithm for character recognition or text recognition.
(3) Implement the proposed algorithm and finish programming and debugging.
(4) Test the proposed algorithm using experimental data. The experimental data should contain more than 10 images in each type. There are two types of images, “longbiao” and book cover.
(5) Thesis writing.

Character recognition is using computer technology to recognize characters automatically. This is an important application field of pattern recognition. People need to deal with a large amount of texts and reports in daily life. Therefore, the purpose of text recognition is to reduce people's labor and improve efficiency. The main task of the graduation design is to recognize the character of important information on the film “Longbiao” and reference books’ cover. I should use OCR technology, and improve the recognition accuracy. According to study digital image processing and MATLAB software, I can understand the processing of text image and the theoretical foundation of the process. The main subject is how to use MATLAB for image filtering, enhancement, text detection and localization. It is hard to separate the character area from background. The key is to propose an efficient algorithm for text recognition.

There are three specifications: complete the programming and debugging of the algorithm. Automatically or semi automatically realize the character recognition algorithm which has image background. Test more than 20 images, and analyze the results. Finally achieve the algorithm.

1.5.2 Meanings

Firstly, it has commercial value. The “Longbiao” is the symbol that the movie will begin. It also means that the advertisements before movie are done. So the advertising company can know that how many times people will watch their ads and so on. Also, it can help to count the numbers of the movies and can recognize the films’ name, type, year, etc. We can understand the theater’s schedule and the audiences’ preferences, and help other researches. The recognition of reference book cover can be used by the company who closely connected with students’ learning. For example, there is a product to assistant students study on the Internet, when the students face a problem, they may take a photo of the book they learned. So the system should recognize the key information like grades, project, title, etc. Then arrange teachers to answer the questions. You can see there are too many companies need this technology.

Secondly, it has research value. At present, there is not too much knowledge about Chinese character recognition. So according to this graduation project, we can know more about Chinese character recognition and partly estimate the development trend. This is also the innovation point. For example, the deep leaning is very popular these days, the recognition of character has reached the stage of handwriting and ignored the font.

Chapter2 Principle of character recognition

2.1 Image Processing

Image processing is using computer to analyze the image. Image processing generally refers to digital image processing. Digital image refers to the use of industrial cameras, scanners and other equipment through the shooting of a large two-dimensional array. The elements of the array called pixels, and the value is called gray value. Image processing techniques generally include image binaryzation, denoising, segmentation, enhancement, restoration, matching, description and recognition, etc. The common system like Cognex system and intelligent system are gradually emerging at present. [4]

21 century is an era full of information. Image as the visual basis of human perception of the world, is an important means of human access to expression and transmission of information. Digital image processing, that is, the computer image processing, its development history is not long. Digital image processing technology is begun from 1920s, when the submarine cable from London to the New York. The United States transported a picture, using the digital compression technology. Firstly, the digital image processing technology can help people understand the world more objectively and accurately. The human visual system can help people get more than three quarters of the information from the outside world, and images, graphics are the carrier of all visual information. Despite the high discriminating power of human can be recognition on 1000 kinds of color, in many cases, image for the human eyes is fuzzy and is not visible. Through image enhancement technology, we can make the fuzzy even invisible image becomes clear and bright.

In the computer system, according to the color and the number of gray, images can be divided into four basic types: binary image, gray image, index image and RGB image. Most image processing software supports these four types of images.

2.1.1 Binary Image

A two-dimensional matrix of the binary image is made up of “1” and “0”. "0" represents black, and "1" represents white. Since the value of each pixel (each element in the matrix) has only two possibilities 0 and 1, so the data type of the binary image in computer is usually 1 binary bit. Binary images are usually used for text and line drawings scan identification and mask image storage.

2.1.2 Gray Image

The range of matrix elements of gray image is usually [0, 255]. So its data type is generally 8-bit unsigned integer, which is often referred to as the 256 gray image. "0" means pure black, "255" means pure white, the middle number from small to large means the transition color from black to white. In some software, gray images can also be represented by double precision data type. Pixel domain is [0, 1]. 0 represents the black, 1 on behalf of the white. The decimal from 0 to 1 means different gray levels. The binary image can be regarded as a special case of the gray image.

2.1.3 Index Image

Index image file structure is much more complex. In addition to storage the two-dimensional matrix of image, it also includes a two-dimensional array called the color index matrix MAP. The size of the map is decided by the image matrix elements range. Such as matrix elements in the range [0, 255], the size of the map matrix is for 256 x 3, represented by MAP= [RGB]. Three elements in each rows of MAP that separately represents the specified values, which are correspond to the colors of red, green and blue monochromatic. Each row of MAP corresponds to image pixel matrix of a gray value, such as a pixel gray value are 64, the pixel is established mapping relationship with the 64 rows of MAP. The pixels on the screen of the actual color are decided by the combination of the 64 rows RGB. That is, when the image is displayed on the screen, the color of each pixel is obtained by searching the gray value of the pixel in the matrix, and the color index matrix MAP is used as an index. The data type of the index image is generally 8-bit

unsigned integer, and the size of the corresponding matrix index map is for 256 x 3. So general index image only shows at the same time with 256 colors. But by changing the index matrix, types of color can be adjusted. The data type of index image can also be double precision floating point type. Index image is used to store the simple color image, such as simple wallpapers in Windows. If the color of the image is much more complicated, we need to use RGB image.

2.1.4 RGB Image

RGB images, same as the index images, can be used to represent the color image. And like index images, it uses red (R), green (G) and blue (B) these primary colors to represent the color of each pixel. But unlike index image, that every pixel RGB image color values placed directly on image matrices. Because each pixel of the color should be presented by composing three components R, G, B. M and N respectively represents the rows and columns. Three M x N dimensional matrixes represent individual pixels of the R, G, and B. The data type of RGB image is generally 8-bit unsigned integer, which is usually used to represent and store the true color images. Of course, it can also store the gray images.

2.2 Pattern Recognition

It is the method that automatically process and interpret the model through computer. We call the environment and the object "mode". With the development of computer technology, it is possible to study the complex information processing process. An important form of information processing is the identification of the environment and the object. For human beings, it is particularly important for the identification of the optical information (obtained from the visual organs) and the acoustic information (obtained from the hearing organ). This is the two important aspects of pattern recognition. The products on the market are optical character recognition, speech recognition system, etc. [5]

Pattern recognition is one of the basic human intelligence. In daily life, people are often in

"pattern recognition". With the appearance of computers in 1940s and the rise of artificial intelligence in the 50's, people certainly hope to use the computer to replace or expand part of human mental work. Pattern recognition developed rapidly in the early 1960s and became a new discipline.

Pattern recognition processes and analyses the character of things or phenomena, then carry on the description, identification, classification and interpretation. It is an important part of information science and artificial intelligence.

Character recognition is one of the applications of pattern recognition. Chinese has thousands of years of history, and is most widely used. So in the growing popularity of information technology and computer technology today, how to input the text convenient and fast to the computer has become an important bottleneck which affects the efficiency of the man-machine interface, but also related to the computer popularity in our country. At present, the Chinese character input is mainly divided into two types: manual keyboard input and machine automatic identification. The manual input is slow and labor intensity is large; automatic input is divided into Chinese character recognition input and speech recognition input. In terms of the difficulty of the recognition technology, the difficulty of handwriting recognition is higher than that of the print character. And in the recognition of handwriting, the difficulty of off-line handwriting is far more than the online handwriting recognition. So far, in addition to the off-line handwritten numeral recognition has been practical application, Chinese characters and other characters of the off-line handwriting recognition is still in the laboratory stage.

2.3 Recognition System

Text recognition generally includes information collection, information analysis and processing, information classification, etc.

(1) Information collection means to transform the gray of the text into electrical signal, and then input to the computer. It is realized by the paper feeding mechanism from character

recognition machine and the photoelectric conversion device such as flying spot scanning, camera, photosensitive element and laser scanning photoelectric conversion device.

(2) Information analysis and processing is to reduce the noise and interference caused by the quality of printing, paper (uniform, stains, etc.) or writing tools. Then does normalization processing such as eliminate, size, deflection, shade, thickness, etc.

(3) Information classification means to classify the results which have removed the noise and normalized. Then output the text recognition results.

2.4 Algorithm of OCR

OCR means using mechanical or electronic to recognize the images of typed, handwritten or printed characters into machine encoded text. It is widely used as a form of data entry from printed paper data records, whether passport documents, invoices, bank statements, computerized receipts, business cards, mail, printouts of static-data, or any suitable documentation. It is a common method of digital printed texts, and it can be electronically edited, searched, stored more compactly, displayed on-line, and used in machine processes such as machine translation, text-to-speech, key data and text mining. OCR is a field of research in pattern recognition, artificial intelligence and computer vision. [6]OCR algorithm is realized by the following four steps. [7-9]

2.4.1 Pre-processing

Process the image which contains text for subsequent feature extraction and learning. The main purpose of this step is to reduce the useless information of the image. Generally, the methods are: binarization, image enhancement, denoising, character segmentation [10], localization [11], normalization, etc. [12-15]

(1) Binarization [16]

It uses the technology of threshold segmentation. It is good at process object with strong contrast from the background. With simple calculation, it can distinguish overlapping regions by closed, connected boundary. If the image pixel gray value greater than or equal to the threshold, it will be judged to be a particular region, with 255 gray values. Otherwise, the pixel dots were excluded and judged to be background or other useless area, with 0 gray values. Printed texts generally have big difference with the background, so it is suitable for setting the threshold for binarization directly.

(2) Image Enhancement

Image enhancement processing can be divided into enhancement based on spatial domain and enhancement based on frequency domain. The step of space denoising is to improve image integrity by reducing pseudograph from acquisition systems. This procedure can reduce the image of small space change. Although the image may be distorted of the original image, region of interest can remain intact because of its strong contrast. The noise displayed by gray-scale map can be seen as the small random changing pixel value relative to the original value.

(3) Denoising

Denoising means using the Gauss smoothing filter to remove noises. It is very important for binarization. It has great influence on the quality of noise reduction algorithm for feature extraction.

(4) Character Segmentation

It is to break up the text in the image into characters. We need to recognize the words one by one. The purpose of Chinese character segmentation is to separate a single Chinese character from the whole image by using the space between the word and word. The characters of Chinese characters are divided into lines and words. Line segmentation is the use of linear space between lines to distinguish the line, line and record the lower and upper bound. Typical algorithms, from top to bottom, accumulate the binary character dot matrix

of pixels per line values. If you start from a row of several additives which are more than a test constant, that the bank is the beginning of a Chinese character, and it is upper bound. Similarly, when a row about a Chinese character height accumulation appeared a series of small accumulation or even zero, it is lower bounds. Word segmentation uses straight lines between word and word space to distinguish the word, and record the left edge and the right edge of the word. Typical algorithms, the upper bound and lower bound of this line is determined, search the text from left to right, cut separate words or punctuation. From the left, began to accumulate row spacing perpendicular to the direction of the pixels. If it is more than a test constant, it can be considered to be the left boundary of the Chinese character. Similarly, when the continuous accumulation of a Chinese character width and suddenly appeared a series of small accumulation or even zero, it can determine the right boundary of the Chinese character. For Chinese text line, the word segmentation is far more difficult than for line segmentation.

(5) Normalization

The normalization is to regularize the text into the same size. It transforms the text size into uniform size, corrects text position and normalizes text stroke thickness, and only on text image projection. Normalization of Chinese characters often brings two problems: first, the scaling of the character image may introduce some interference; the second is the image scale itself. Therefore, it is necessary to use the appropriate normalization method to eliminate the influence of scale change on the feature extraction. We can only use the same specifications under the unified algorithm. If the text is tilt, it also needs tilt correction.

2.4.2 Feature Extraction and Dimensionality Reduction

Features are the key used to identify the characters. Each different character can be distinguished by the features from others. For numbers and English letters, feature extraction is relatively easy, because the numbers are only 10 and English letters are only 52. They are small character sets. For Chinese characters, feature extraction is much more difficult, because the Chinese character is in a large character set. In the national standard,

the most commonly used Chinese characters are 3755. As well as complex structure of Chinese characters, there are many characters with similar form. After determined which feature to use, depending on the circumstances, there may be a need to reduce the dimension of feature. If the dimensionality of the feature is too high (generally feature present as a vector, the dimension is the vector component number), the efficiency of the classifier will have great influence. In order to improve the recognition rate, we need to reduce the dimension. This process is also very important. It is not only reduces the dimension but also retains the enough amount of information to distinguish different words.

2.4.3 Classifier Design, Training and Practical Recognition

Classifier is used for recognition. For a text image, we need to extract the features and throw them to the classifier for classification. It will tell you the feature should be recognized to which word. Before the actual recognition, we often train the classifier which is a supervised learning case. There are many mature classifiers such as SVM, kn, neural network, etc. Character recognition method basically divided into statistics. Methods commonly used are template matching and geometric feature extraction. [17-19]

(1) Template Matched

Template matching means to match the input text with a given standard text or template, calculate the similarity between them and take maximum category as a result. The disadvantage of this method is when the number to be identified increases, the number of standard text or templates also increased. This will increase the storage capacity of the machine; on the other hand will reduce the recognition accuracy. So this method is suitable for the printed text identification which is fixed font. The advantage of this method is to calculate the similarity with the whole text, so it has strong adaptability for the defect and edge noise.

(2) Geometric Feature Extraction

Geometric feature extraction means to extract some geometric features of characters, such

as word’s endpoint and bifurcation points, concave and convex portions and line in each direction like horizontal, vertical and inclined, closed loop, and so on. According to these features’ location and relationship, we can do logic judgment, and obtain the identification result. This method is use structural information, so it is suitable for handwritten text as larger variant.

2.4.4 Post-processing

Post-processing is used to optimize the results. Firstly, classification sometimes is not completely correct (actually can not entirely correct), such as the recognition of Chinese characters. Due to characters with similar form, it is easily recognize the word into a similar one. Post-processing can solve this problem. For example, use the language model for correction. Secondly, the image is often containing a lot of text, and the text exists in complex situation. Post-processing can try to format the recognition results, such as arrange the type setting.

2.5 Accuracy

Commissioned by the U.S. Department of Energy (DOE), the Information Science Research Institute (ISRI) had the mission to foster the improvement of automated technologies for understanding machine printed documents, and it conducted the most authoritative Annual Test of OCR Accuracy from 1992 to 1996. Recognition of English and numbers is still not 100% accurate even clear imaging is available. One study based on recognition of 19th and early 20th-century newspaper pages concluded that OCR accuracy for commercial OCR software varied from 81% to 99%. There are few results about Chinese character recognition. The MNIST database is commonly used for testing systems' ability to recognize handwritten digits.

Chapter 3 Recognition of “Longbiao”

3.1 Introduction of “Longbiao”

“Longbiao”（Fig.3-1）is the leading logo before the movie showed on the screen. When you watch a movie in the cinema, you can always see it. It is the license for public projection. It is issued by the State Administration of Radio, Film & Television （SARFT）in China. From the picture of “Longbiao”, we can know the film type, approval label, year and other important information. Therefore, we need to get information like that.

Fig.3-1 An Image of “LongBiao”

“电审故字”, represents the film type, which is issued by Film Bureau, State Administration of Radio, Film & Television. “电审” means the films are reviewed by the SARFT. Of course, the “电审数字” refers to the pure digital films. “电审进字” refers to the foreign films that pass through the review, including the introduced 2D split tablets and tablets group. “电审（特）字进字” refers to the special introduced movie that is 3D stereo movies which we prefer to watch in the cinema. “电审动字” refers to the domestic animation film; “电审故（复）字” refers to some of the repaired old pieces for release,

such as “New Dragon Inn”, “ghost story”.

From state regulations, the film must pass through the review and get a permit for public screening (that is “Longbiao”). Then it will be allowed for showing in the public. The domestically produced films must get the release permit (“Longbiao”) then allowed the release of projection and participate in some international film festival or competition.

3.2 Recognition

3.2.1 Image Features

The background of this kind of image is single, and the key information is fixed. The critical information to be identified is:“电审 X 字 [XXXX (year)] 第 X 号” below the “公映许可证”. Other texts will not change, so they do not need to be specifically identified. The image resolutions are different. The images are separated into two types: Original and Remake as shown in Fig. 3-2.

Fig. 3-2(a) Original

Fig. 3-2(b) Remake

The recognition flowchart is shown below (Fig.3-3):

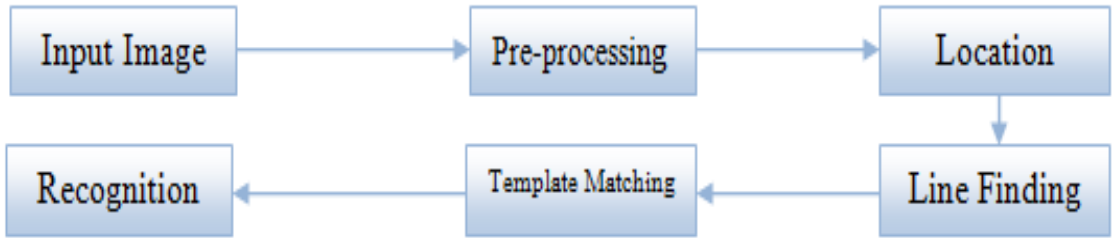

Fig. 3-3 The flowchart of recognition

3.2.2 Pre-processing

Firstly, convert RGB true color image to gray image. In this situation, we do not need the binarization, because the effect of binarization is not good. The result after pre-processing is shown below (Fig.3-4):

Fig.3-4 The result after pre-processing

3.2.3 Text Location

For normal recognition, I saw the result. Some important information will be identified wrong. For example, there are mistakes between “5”and “S”, “0”and “o”,“1”and “L”, “|”, etc. Figure 2 shows the confusion about “5”and “S”, “1”and “|”.

The prior knowledge is useful, so I can use text location. The words in the middle of the “[ ]”must be digital, which present the year. The words behind “第” are also digitals, which present the number. So let information on these locations only compare with the digital library. We can use the “CharacterBoundingBoxes” to determine the location of the characters. Set the parameters “characterset” to be “0123456789” in the OCR functions. This method can rule out some of the impact from other words with the similar shape.

Fig. 3-5 Text location

3.2.4 Line Finding and Text Segmentation

The line finding algorithm is used so that an inclined page can be identified without having to de-skew, and saving the loss of image quality. The important part of this process is construction the line and character filtering. Assume that recognition analysis has already segment the text into a roughly uniform text size, a height filter removes drop-caps and vertically characters. The average height is roughly the character size in the region, so it is safe to filter out blobs that are smaller than the average height, which is most likely punctuation, marks and noise, or something useless and influence the result. The filtered blobs are more likely to adapt a non-overlapping and parallel model, but with a sloping lines. Sorting and processing the blobs by x-coordinate can make it possible to deliver the blobs to a unique line. While tracking the slope line across the page, the danger of assign an incorrect text of tilt is reduced. Once the filtered blobs have been assigned to lines, an average of squares fit is used to estimate the baselines, and the inclined blobs will be fitted back into the correct lines. The last step of the line creation process is that overlap by at least half horizontally. Then put diacritical marks together with the correct standards and will finally collect parts of some broken characters. Then use threshold segmentation algorithm to do the text segmentation.

3.2.5 Template Matching

Match the characters with Chinese character library. So I should find suitable Chinese character library at first. Then compare with the character template one by one. The method is let the recognized character subtract with the template. Then calculate the value of all the

pixels after subtraction. If it is less than a threshold, the recognized character and the template is the same character. Thus complete one of the recognitions. Then loop to identify the characters with the same processing, and can identify all the characters. The results are stored in the string.

3.3 Results

The program can recognize the key information basically. The accuracy is greatly improved when add text location. The results are as follows:

Example of Original Image (Fig. 3-6):

Before location:

Fig. 3-6(a) Original Before

Fig. 3-6(b) Original Before (Detail)

The result has some wrong results. It can't recognize the specific number of the movie. Because there are too many form near words.

After location:

Fig. 3-6(c) Original After

Fig. 3-6(d) Original After (Detail)

The final result should be "电审故字【2013】第157号". Errors of recognition cannot be avoided, but it does not affect the identification for the key information. Although there still has some mistakes, the key information is correct. The final character will always be "号", so the result is reasonable. I don't need to continue processing it.

Example of Remake Image (Fig. 3-7):

Before location:

Fig. 3-7(a) Remake Before

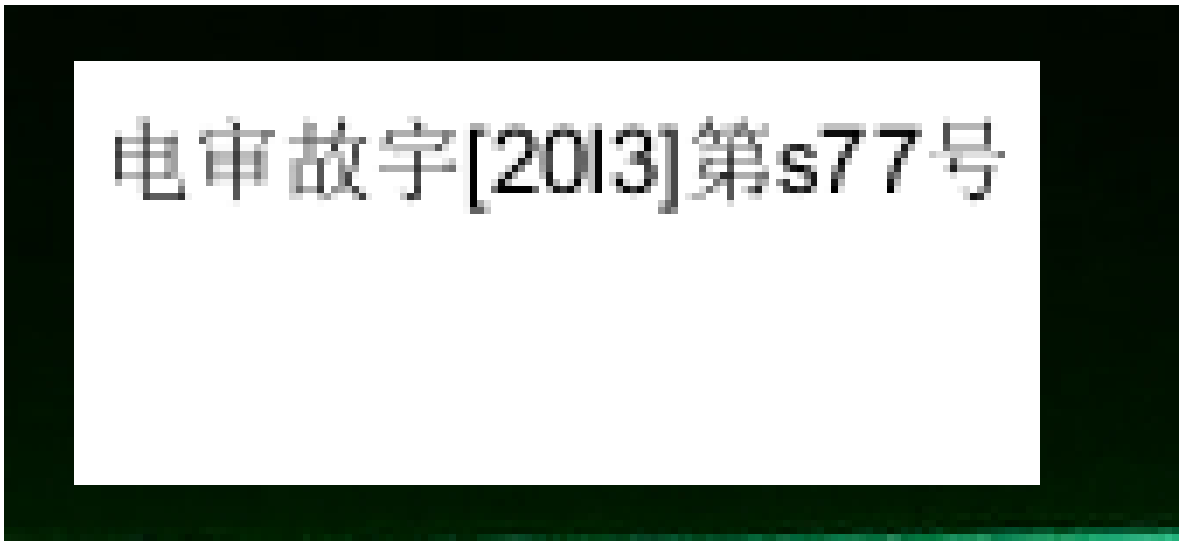

Fig. 3-7(b) Remake Before (Detail)

The key information like year and serial number are all wrong. They should be numbers.

After location:

Fig. 3-7(c) Remake After

Fig. 3-7(d) Remake After (Detail)

The final result should be “电审故字【2013】第 577 号”. You can see all the important information are recognized correctly.

Chapter 4 Recognition of Reference Book Cover

4.1 Introduction of Reference Book Cover

We all use reference books in our student days. Its cover (Fig. 4-1) concludes subject, grade, press, volumes and other important information. So we need recognize the characters on the cover. And we will know which book the student used.

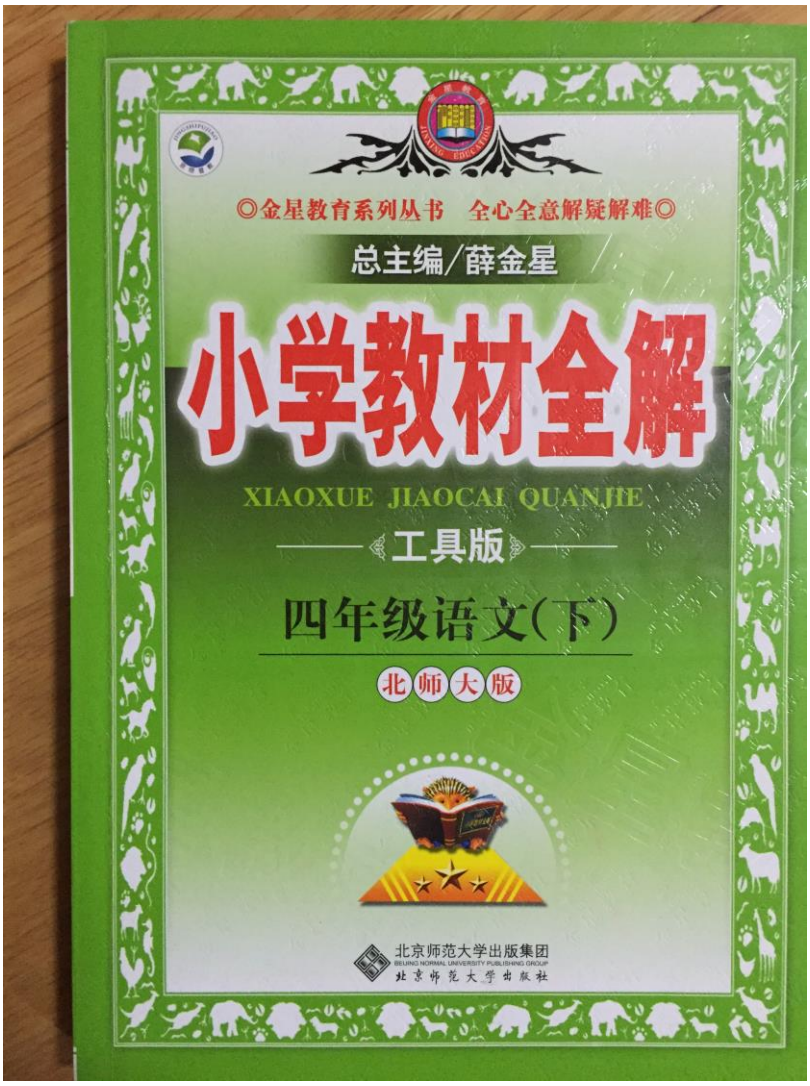

Fig. 4-1 Reference book cover

The images above all need character recognition on an image with background. They mainly need to recognize the Chinese characters and numbers which are all printed.

4.2 Recognition

4.2.1 Image Features

The background of reference books' cover is mixed and disorderly. The text size, font, color are different. It is hard to recognize directly, such as “北师大版” located on the white circle background. It needs background processing. The key information to be identified is: the title (such as “小学教材全解”), grades (such as “四年级”), subjects (such as “语文”), edition (such as “北师大版”), etc. It needs to recognize Chinese characters which are all printed. Images are taken by phone camera, so the resolution is not high.

The recognition method and flow chart are almost same with the recognition of “Longbiao”. But there still has some differences. The book cover has complex backgrounds, which make the recognition much more difficult. So it needs the background processing using RGB.

4.2.2 Background Processing Using RGB Channel.

This is a part of preprocessing. A RGB image is an MxNx3 array of color pixels. Every color pixel points a specific location of color image that corresponds to the red, green and blue. There will be a range of values presents every color. The color of the words to be recognized is same. So firstly select the area which should be recognized and get the RGB value of the text color. Then use and-logic and non-logic in the image matrix to assign text to black. And other areas are backgrounds that assign to white. This method realizes the background assimilation and binaryzation. The areas processed are as follows (Fig. 4-2, Fig. 4-3):

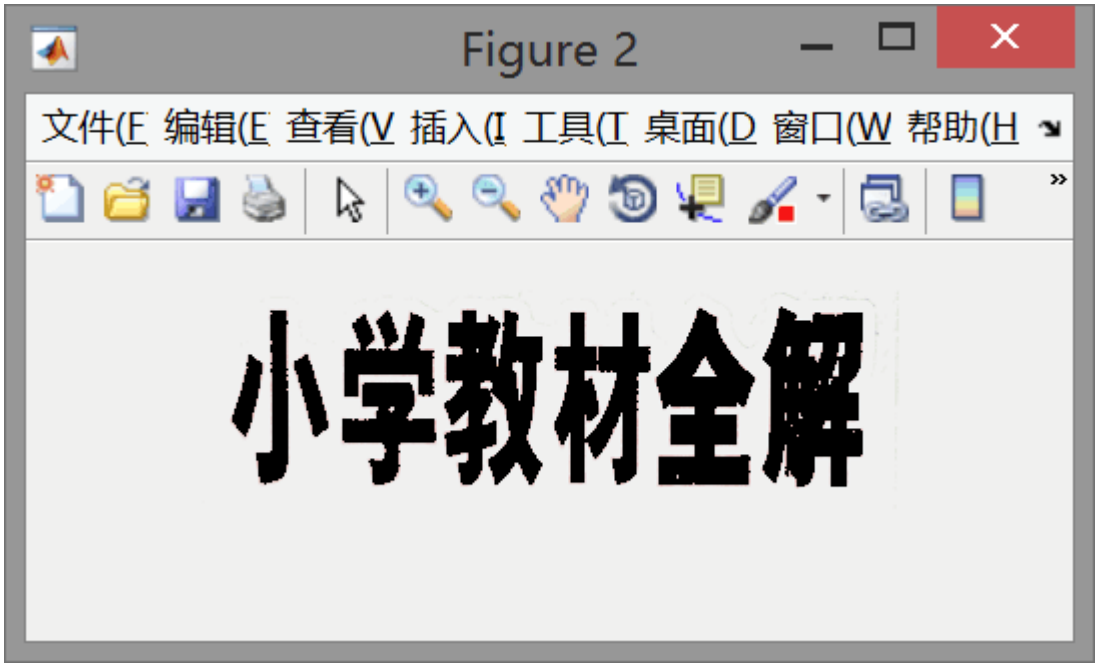

Fig. 4-2 The Title Information After Background Processing

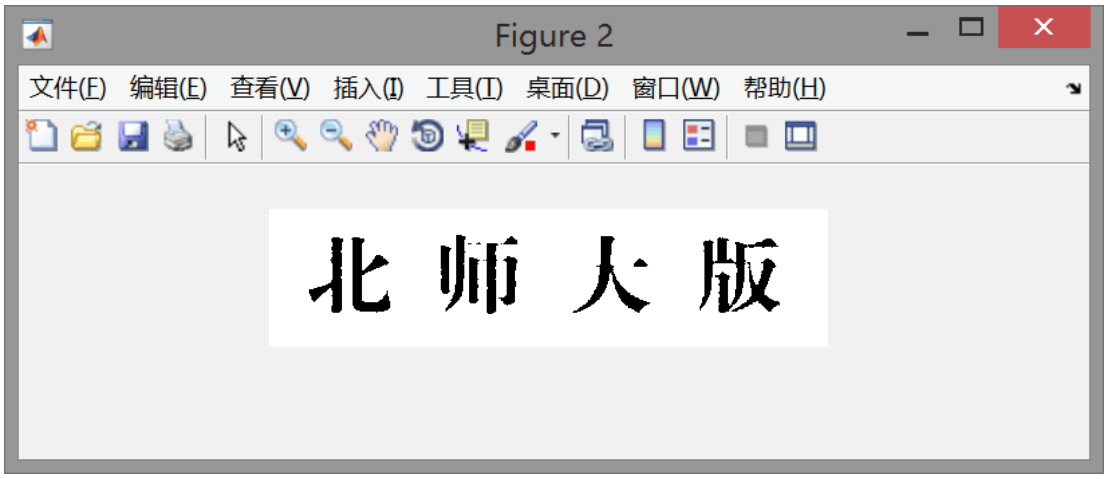

Fig. 4-3 The Edition Information After Background Processing

The result shows it just selected the texts. There is no influence on background.

4.3 Results

The information about grade and subject do not need background process, because the backgrounds are simple. Fig. 4-4, 4-5, 4-6 show the recognition results of all key information.

Fig. 4-4 Recognition Results of the Grade and subject

The result is correct.

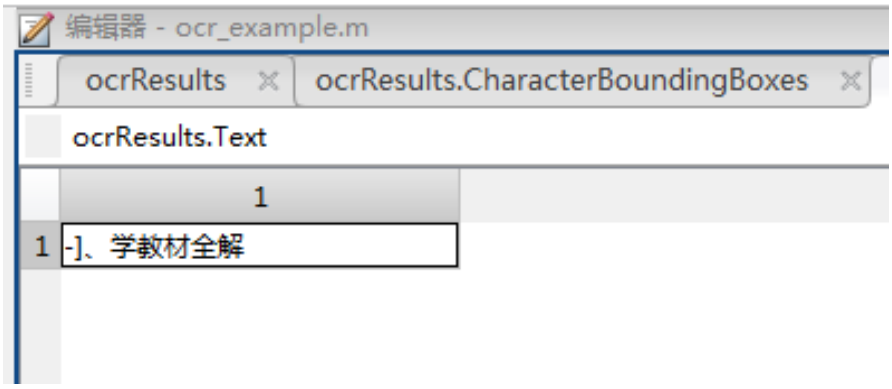

Fig.4-5 Recognition Result of the Title

There is a mistake about “小”. The Chinese character library is limited, only contains a few fonts. This character are showed in wordart. So it is hard to be recognized.

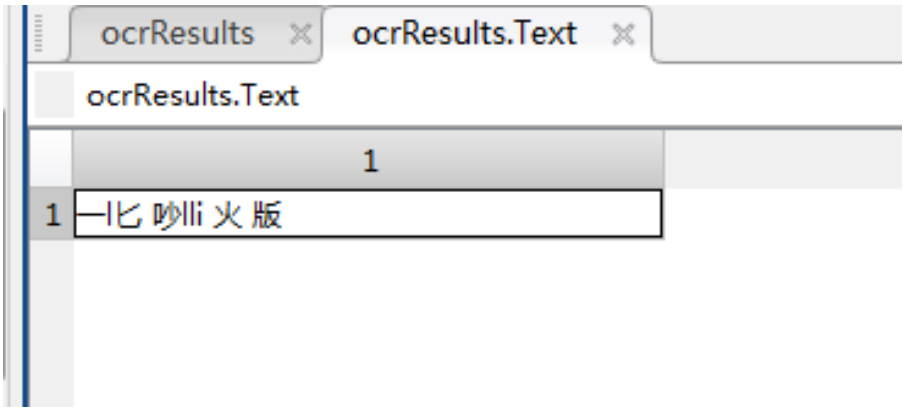

Fig. 4-6 Recognition Result of the Edition

This result is bad. Because the image pixel is too low, and there are influences like reflection and different fonts. After pre-prosessed, part of the character will be erased. So the character “大” may looks like “火”, etc. OCR has its own limitations. It is unable to achieve the accuracy rate 100%. We still need work on it.

4.4 Difficulties in Chinese Character Recognition

This graduation project includes the recognition of Chinese characters. Chinese character recognition is much more difficult than English and number in the following aspects:

4.4.1 Too Many Categories

The commonly used Chinese characters are about 3000~4000. GB2312-80 has 6763 commonly used Chinese characters, divided into two grades. First grade has 3755 Chinese characters, the using frequency is 99.7%; the second grade has 3008 characters. Recognition system should generally correctly identify all of these common words in order to meet the needs of practical application. The current research aims at recognize the 3755 Chinese characters, even it is, there are a large numbers of Chinese characters to be identified. It can

be said that the large amount of word is one of the main reasons for the difficulty of Chinese character recognition.

4.4.2 Complex Structure of Chinese Characters

Chinese is a strong structural character. Each character has unique and specific strokes. Stroke is the most basic part of Chinese characters. Chinese characters can be seen as a combination of components. Component is the combination of strokes. It is commonly known as radicals. The strokes and radicals of different combinations form thousands of different meanings of abnormal complex Chinese characters. Compared with other nationalities in the world, the structure of Chinese is the most complicated character.

4.4.3 Many Similar Characters

There are only very small differences between each Chinese character. With the same stroke, there are small changes in a stroke position or form. For example, the "土" and "士" . This two words only have subtle differences in the lower part of the stroke length. Even though these Chinese characters recognized by the human eyes, it is easily confused. Identification algorithms and systems must be able to correctly these subtle differences, otherwise it will be wrong. There is no single feature can complete recognition of Chinese characters, so how to effective select the features, how to effectively combine, how to get a quickly matching speed and high recognition rate are the big problems in Chinese character recognition.

Chapter 5 Conclusion and Thoughts

5.1 Recently Achievement

(1) Studied the characteristics of "Longbiao" and cover of reference books;

(2) Understood the location of the identified key information;

(3) Mastered the method of OCR;

(4) Located the key information of "Longbiao";

(5) Basically realized the automatic character recognition of the information on "Longbiao";

(6) Preprocessed with RGB channel on reference books cover;

(7) Basically realized the character recognition on the cover;

(8) Tested more than 20 images for each kind of images.

5.2 Existing Problems

(1) The accuracy rate of book covers is not high, because the Chinese characters library used for template matching may be not perfect.

(2) The photos are taken by phone camera, so they may have low pixels and reflection, which influence the recognition.

(3) The key information on the cover is scattered, so it is temporarily unable to automatically identify. Now it needs select the recognition range manually. Or it will influence the accuracy.

(4) OCR has its own limitations. It is unable to achieve the accuracy rate 100%, especially in Chinese character recognition. There is a table of error rates over various languages in others thesis (Table 5-1). We can see the error rates in Chinese recognition are very high, even no data. [20]

Table 5-1 Error rates over various languages

Language	No. of chars (millions)	No. of words (millions)	Char error rate (%)	Word error rate (%)
English	39	4	0.5	3.72
EFIGSD	213	26	0.75	5.78
Russian	38	5	1.35	5.48
Simplified Chinese	0.25	NA	3.77	NA
Hindi	1.4	0.33	15.41	69.44

5.3 Outlook

With the further development of society, character recognition has important practical value and theoretical significance in information processing, office automation, postal system and financial system. In addition, in our five thousand years of history, history has left us many valuable documents and records which are also need to be stored in the database. So the development of Chinese character recognition is much more important. Overall, in recent years, the research on the character recognition still needs deeply study. We have also achieved some great achievements, so that the recognition rate of the system is rising. The application of Chinese character recognition system will be more and more mature.

A group of outstanding commercial OCR software, such as Hanwang OCR, Tsinghua TH-OCR2000 has been created. Current recognition methods have begun to research on the identification and post-processing, combined with the semantic understanding of after treatment before recognition and pre-processing should be improve the recognition correct rate. According to the analysis of human character recognition, the text is generally in the context of a combination of understanding. Therefore, computer in the character recognition should also base on combined contextual information on the recognition results correction. The recognition result will be a complete word or a sentence, which based on the statistical information of the language.

Recognition algorithm and processing method can improve the recognition rate and reduce the error recognition ratio, weaknesses, and mutual compensation to form optimal combination. Along with the continuous optimization of the identification method and the continuous maturation of the post-processing technology, the combination of the two methods will become the research direction of Chinese character recognition in the future. I think that character recognition technologies in the future will focus on breakthroughs in the following aspects:

5.3.1 Chinese Character Recognition Rate

It is always one of the most important indicators of Chinese character recognition. It should reach the height that greatest reduce user workload about proofreading.

5.3.2 The Automatic Analysis of Layout

Without artificial intervention, we can recognize the printed text materials, such as newspapers, magazines, above all illustrations, tables, lace and exist on both vertical and horizontal layout. And it will be easily to be distinguished and marked.

5.3.3 Further Improvement of Overall Performance

Solve the problems like automatic understanding some kind of complex newspaper columns and position arrangement; using natural language understanding knowledge to do post-processing; further improve the character recognition rate and adaptability; decrease recognition error rate, and so on.

5.3.4 Self-learning Function

So that users can freely expand the professional recognition character set, and can keep pace with the times, suitable for a variety of application environment of Chinese character

recognition system. For example: Mac OS, Windows and UNIX. It can meet the needs of different users.

5.3.5 Online Version

Make full use of the resources and computing power of the network to improve the performance of the algorithm, so that users can work together more easily.

5.3.6 Deep Learning

Let’s talk about the deep learning. The concept of deep learning is based on artificial neural networks. Multilayer perceptron with multiple hidden layers is a kind of deep learning structure. Deep learning by combining low-level features to form a more abstract high-level representation of attribute categories or features in order to discover the distributed feature representation of data. [21]

The concept of deep learning was proposed by Hinton, 2006. Based on the depth confidence network (DBN), an unsupervised greedy algorithm is proposed to solve the problem of deep structure related optimization. The convolution neural network was proposed by Lecun. It is the first real multi layer structure learning algorithm, which uses the relative relation of the space to reduce the number of parameters to improve the training performance.

Deep learning is a new field in the research of machine learning. The motivation is established that is to simulate the human brain to analyze learning of the neural network, which mimics the mechanisms in the brain to interpret the data, such as images, sound and text. With machine learning methods, depth of machine learning methods have the unsupervised learning and supervised learning. Different learning framework for establishing the learning model is very different. For example, convolution neural network (CNNs) is a depth of supervised learning machine learning models and deep belief network (DBNs) is a non supervised learning machine learning model. [22-23]

5.4 Thoughts

From the end of January, I began to do the graduation design. After several months hard working, I finished my graduation design, which can accurately recognize the text information on some specific pictures. During this period of time, I have mastered the ability of MATLAB programming, and deepen the understanding of the character recognition algorithm. More importantly, my ability of retrieve the literature has been further improved. Prior to the character recognition processing, my MATLAB knowledge was not enough, I need to reading a lot of literature not only to enrich my background, but also to expand my horizons. When facing the new knowledge in the future, I can quickly grasp the effective retrieval of useful information. This ability is much more valuable in the future study and life.

At the same time, under the supervision of Prof. Li Huiqi, I did not drag, and progress every week. Overall, my algorithm is ideally achieved the desired recognition results, but there is still a lot of room for improvement. Character recognition is not only the simple print or the recognition of a single language, through the deep learning; it can even recognize more complex fonts such as handwriting. I collected information about deep learning of character recognition such as caffe-ocr when I finished my graduation project. I configured running environment well, and see the previous results, could not help but sigh, deep learning is the field that worth us to further explore.

The graduation design is a valuable way from the theoretical design into the practical demonstration. It can be said that it is assessment and summary result for my four-year university study. Looking back at this period of time, although I have encountered many hardships in the process of learning, I learned a lot. Character recognition is a new field for me, but under the guidance of teachers and students in few months, I have mastered the basic principles and design methods of this technology. And I realized simple Chinese character recognition, which is using MATLAB programming. The document editing and

typesetting has also been considerable progress. The benefit bandit is shallow. Not only that, through this graduation project I also enhance my thinking ability, communication skills, and many other abilities that textbooks cannot tell, which also allows me to improve my comprehensive ability. Of course, although I successfully finished the graduation project, I know that my knowledge is far from enough. There are also some errors and shortcomings. Character recognition is this field that waiting me to deeply explore. In the future study, I will always work hard.

Acknowledgements

This paper comes to the end; my university life also comes to an end. At this time, I would like to thank the people who helped me a lot.

First of all, I would like to thank my advisor, Prof. Li Huiqi. She is strict, but also kind; she pursuit efficiency and order, but also courteous and accessible. It is she that taught me to do things in schedule and because of this, do not delay and drag. Every week in the lab meeting, she treats us equally, whether you are an undergraduate or graduate student, you can get her guidance in detail, which makes my graduation project very smoothly. Thank you, Prof. Li!

Secondly, I would like to thank the seniors in the laboratory. Wang Shumeng guided me most closed. She has a lot of experiences. Every time I have a problem, she helps me selfless. And our communication is very smooth. There are also other seniors not only helps me a lot but also let me feel the laboratory like a big family. We hang out together, barbecued, played games together. They make my life more colorful, and leave me too many beautiful and moving memories.

I would also like to thank the students together to do the graduation project. We can work together to discuss the progress, the future plans and arrangements. We helped each other, and they gave me too much encouragement and support.

In addition, I would also like to thank all the teachers and students I met during the University. You all witness my growth; you bring me warmth; you are the content of my college life. Graduation project is the end of the University, but also is the beginning of a new journey. I will go with the power and moved you gave me, and go on without hesitation.

Reference

[1] Clavelli A, Karatzas D, Llados J. A framework for the assessment of text extraction algorithms on complex colour images[C]. Iapr International Workshop on Document Analysis Systems. 2010:19-26.

[2] 齐爱军. 浅谈文字识别软件 OCR[J]. 印刷技术, 2004(13):27-30.

[3] 田浩鹏, 董怡彤. 关于数字图像处理技术的研究[J]. 北方经贸, 2010(12):140-141.

[4] Gonzalez R C, Woods R E, Eddins S L. Digital image processing using MATLAB[M]. Digital image processing. Pearson Prentice Hall, 2007:197-199.

[5] 史海成, 王春艳, 张媛媛. 浅谈模式识别[J]. 今日科苑, 2007(22):169-169.

[6] 朱志刚等译. 数字图像处理[M].北京:电子工业出版社,1998.9.

[7] Smith R, Antonova D, Lee D S. Adapting the Tesseract Open Source OCR Engine for Multilingual OCR[M]. Adapting the Tesseract open source OCR engine for multilingual OCR. 2009:1-8.

[8] Smith R. Hybrid Page Layout Analysis via Tab-Stop Detection[C]. Proceedings of the 2009 10th International Conference on Document Analysis and Recognition. IEEE Computer Society, 2009:241-245.

[9] Lowe D G. Distinctive Image Features from Scale-Invariant Keypoints[J]. International Journal of Computer Vision, 2004, 60(60):91-110.

[10] Wolf C, Jolion J M. Object count/area graphs for the evaluation of object detection and segmentation algorithms[J]. International Journal of Document Analysis & Recognition, 2006, 8(4):280-296.

[11] Zhong Y, Jain A K. Object localization using color, texture and shape[M]. Energy Minimization Methods in Computer Vision and Pattern Recognition. Springer Berlin

Heidelberg, 1997:671-684.

[12] Kai W, Babenko B, Belongie S. End-to-end scene text recognition[C]. 2011:1457-1464.

[13] Everingham M, Eslami S M A, Gool L V, et al. The Pascal, Visual Object Classes Challenge: A Retrospective[J]. International Journal of Computer Vision, 2014, 111(1):98-136.

[14] Lindeberg T. Edge Detection and Ridge Detection With Automatic Scale Selection[J]. International Journal of Computer Vision, 1996, 30(2):465-470.

[15] Antani S, Kasturi R, Jain R. A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video[J]. Revista Brasileira De Ciência Do Solo, 2010, 35(4):945-965.

[16] Ntirogiannis K, Gatos B, Pratikakis I. An Objective Evaluation Methodology for Document Image Binarization Techniques[C]. The Eighth Iapr Workshop on Document Analysis Systems. 2008:217-224.

[17] Flickner M, Sawhney H, Niblack W, et al. Query by Image and Video Content : The QBIC System[J]. Computer, 1995, 28(9):23-32.

[18] Bishop C. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer[J]. Journal of Electronic Imaging, 2006, 16(4):140-155.

[19] Garcia C, Apostolidis X. Text detection and segmentation in complex color images[C]. International Conference on Acoustics. IEEE, 2000:2326-2329 vol.4.

[20] 盛利元，李宏言，孙克辉．“数字图像处理”实验教学探索与实验软件研制［J］．电气电子教学学报, 2005, 27(3):75-77.

[21] 孙志军，薛磊，许阳明, 等．深度学习研究综述[J]．计算机应用研究，2012,

29(8):2806-2810.

[22] Kim H K. Efficient Automatic Text Location Method and Content-Based Indexing and Structuring of Video Database[J]. Journal of Visual Communication & Image Representation, 1996, 7(4):336-344.

[23] Smith R. An Overview of the Tesseract OCR Engine[C]. International Conference on Document Analysis & Recognition. 2007:629-633.