OCR RELATED TECHNOLOGY TRENDS

Seung Ju Jang
College of ICT Engineering, Dong-Eui University
Korea
sjjang@deu.ac.kr

ABSTRACT

The technology related to character recognition has emerged as an important technology in the era of the fourth industrial revolution. Character recognition is developing as a core technology needed in various fields. Character recognition is performed by extracting characters from an image and recognizing the extracted characters. Character recognition technology has been continuously developed. Recently, along with the development of the fourth industrial revolution, character recognition technology has been used as a core technology in many places. This paper introduces the technology related to character recognition and the program for character recognition.

Keywords: OCR, OCR Program, OCR Technology, OCR Technology Trend

1. INTRODUCTION

In the era of the 4th industrial revolution, IT technologies are being combined and converged in various fields. Among them, the character recognition field using image information has been utilized in various fields. Optical Character Recognition (OCR) refers to the acquisition of images of characters written by humans or printed by a machine with an image scanner to convert them into machine-readable characters.

Character recognition technology has been continuously researched and developed in the field of OCR. Various kinds of OCR related programs are developed and used. Some OCR programs are open to the public for free, while others are available in paid versions. OCR programs generally aim at extracting text from image data. When extracting text from image data, we could encounter various technical problems. The big problem is that the recognition rate can vary greatly depending on the user's language. This is often caused by the specificity of the language.

In recent years, OCR programs that solve these technical problems more actively have emerged. With the development of these technologies, the use of OCR programs has recently become popular. Document analysis and string extraction from natural images are the most basic and important issues for understanding documents and images. While text recognition is already available in many commercial products, analyzing and recognizing complex documents and natural images is not easy to solve yet. Finding and analyzing characters in various environments and extracting text characters accurately has become an important technology.

Although various OCR-related studies have been conducted, there is a lack of papers on technologies related to character recognition. Therefore, this paper introduces the related technologies and functions required for OCR and summarizes the characteristics of each technology. In addition, I introduce OCR SW program.

2. HISTORY OF OCR TECHNOLOGY DEVELOPMENT

The history of OCR technology development began in 1928. It is a character recognition method using pattern matching. It compares several standard pattern characters and input characters prepared in advance and selects the most similar to the standard pattern character as the corresponding character. Around 1955, devices for recognizing printed numbers were invented in the United Kingdom and the United States.

In the 1960s, various researches were conducted around IBM in the United States. The research on the improvement of the character recognition rate, the handwritten recognition problem, etc. has been conducted. In the third generation OCR system, which appeared in the mid-1970s, the quality of documents for character recognition was poor. The environment for OCR was primarily used as a tool for simple typing before the advent of PCs.

As the computer hardware performance improved in 1980, software development for OCR progressed, and commercial systems began to spread. Due to the proliferation of computers in the 1980s, the requirements for character recognition began to increase. At this time, they actively conducted researches on character recognition in the culture of Chinese characters. Recently, due to the spread of smartphones, research on the field of character recognition using smartphones has been actively conducted.

The study of technology related to character recognition is conducted mainly in English language. This is because English-related documents are mainstream in our life. Recently, they expanded researches to study the character recognition of various languages around the world [1, 2, 3].

3. OCR RELATED TECHNOLOGY AND APPLICATION FIELD

Techniques related to character recognition are required in various fields. In particular, character recognition technology using image processing technology is rapidly developing. While these technologies are developing rapidly, there are many technologies that need to be addressed in computer recognition.

The process of character recognition is generally divided into three stages: preprocessing, region analysis, and recognition. The process of preprocessing for character recognition is an important part to increase the character recognition rate. In the program for character recognition, character recognition is performed using data that has been preprocessed. Character recognition also differs depending on language. This paper focuses on print English character, which is commonly used.

The preprocessing process for character recognition is as follows. For character recognition, the extraction work on the text area must be preceded. The extraction on the text area for character recognition is performed in various ways. The image for character recognition is converted into a form that is recognizable through preprocessing. In this process, unnecessary information is removed and conversion for binarization is performed. In particular, the handwriting character is often informal, and thus, a tilt correction process is necessary [2]. A typical character recognition process is shown in Fig. 1 below.

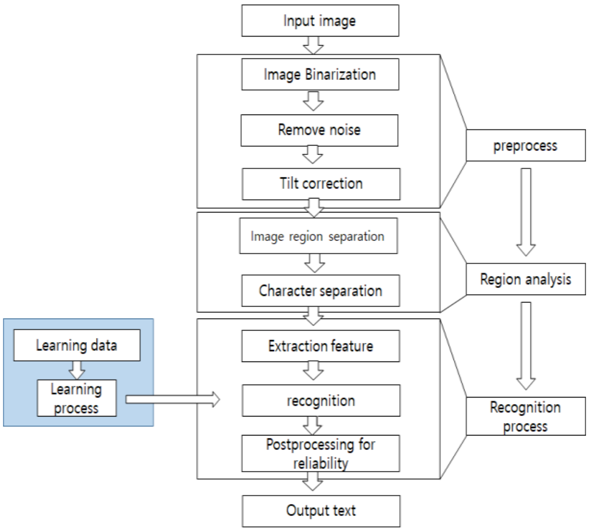

Figure. 1. Character Recognition Process

Character reading can be divided into two parts: preprocessing of input data and character recognition based on processed data. Preprocessing procedure is as in Fig. 1.

Image data goes through the binarization module, thinning module and straight line extraction module. Image data are processed and have a similar format, and then the data is suitable for character recognition through the gradient difference module, the thinned line separation module, the Bezier curve calculation module, and the pen thickness calculation module. Then, the characters are recognized through the character attribute graph generation module, the basic stroke graph matching module, and the character graph matching module [2].

Before processing the binarization process, we first go through a gray scale transformation for contour extraction. Next, the boundary of the text area for the text recognition image is extracted. Various operators are used for contour extraction.

Binarization refers to converting an input image into a binary image of 1 or 0. The reason for binarization is to distinguish the background from the characters clearly. In binarization, the characters can be extracted accurately according to the threshold setting for distinguishing the characters from the background. There are various ways to find the threshold.

Image segmentation is one of the ways commonly used to classify the pixels of the image correctly in the video image. This divides the image into several distinct regions so that the similarity of pixels in each region is high and the contrast ratio between regions is high.

Image segmentation is one of the basic problems of image analysis. The importance and usefulness of image segmentation allows us to analyze images accurately using extensive research and several proposed approaches such as intensity, color, and texture. There are various image segmentation techniques based on threshold, boundary, cluster, and neural network [2, 3, 4]. An important part of preprocessing for character recognition is removing noise. Noise is an unwanted pixel deformation of an image that reduces the effectiveness of the image processing mechanism. Gaussian Blurring Smoothing technology is used to remove small amounts of noise.

Region analysis procedure is as in Fig. 1. For image segmentation, clustering algorithms such as Fuzzy C-Means have been developed. Image segmentation identify cluster prototypes as dots in the image partition and determine the membership function of each pixel. This typically divides the image into regions that are homogeneous in some sense or centered around regions of significance. Other algorithms for segmentation like this include K-Means clustering, expectation maximization algorithm, and mean shift algorithm [4, 5].

Through the process of session development, it extracts the string from the image. It extracts the characters that correspond to the string and converts them into characters. Character feature extraction is mainly used for handwritten character recognition. Feature extraction is processed based on original image information such as color, structure, texture, projection histogram, and instantaneous invariant. Once feature extraction is complete, the classifier is used to classify the characters. Commonly used character classification methods mainly use template matching, K-Nearest Neighbor algorithm, artificial neural network algorithm, support vector machine, etc. [5].

Image recognition and pattern recognition among character recognition applications can identify various different cases of images or patterns. Image recognition technology focuses on classifying and extracting images. Pattern recognition deals with recognizing patterns in images or data sets. OCR is a very important technology that can be applied to various fields in the 4th industrial revolution because it enables text recognition from a given image. OCR technology even makes it possible to identify characters in handwritten scripts [4].

3. 1. Character Recognition

Character recognition process is as in Fig.1, Fig.3. Handwritten character recognition has a great potential for application, and there is a great demand in accordance with the development of society in industries such as an image recognition system or a handwriting input device. Many researchers are interested in the study of handwritten character recognition in the industry. In particular, in the field of image processing and pattern classification, handwritten character recognition has been extensively researched and developed. The methods currently used for handwriting recognition fall into two categories: handwritten character recognition based on pattern classification and in-depth learning as in Fig. 2.Character recognition is a process of extracting features of various images. One of the important technologies, Surface Mount Technology(SMT), or Back Propagation Neural Network(BPNN) is used to handle character recognition [6].

Character recognition can find solutions through approximate optimal solutions such as heuristics algorithms due to various difficult problems. In general, the execution time is often an exponential function in the character recognition algorithm. Therefore, they have conducted researches on character recognition methods using various methods. Some cases used Genetic Algorithms as a probabilistic approach.

Optical Character Recognition is mainly focused on text recognition for printed documents created using a text editor. Convolutional Neural Networks (CNN) are increasing the accuracy of the computer vision and pattern recognition community, especially offline handwriting recognition technology [8].

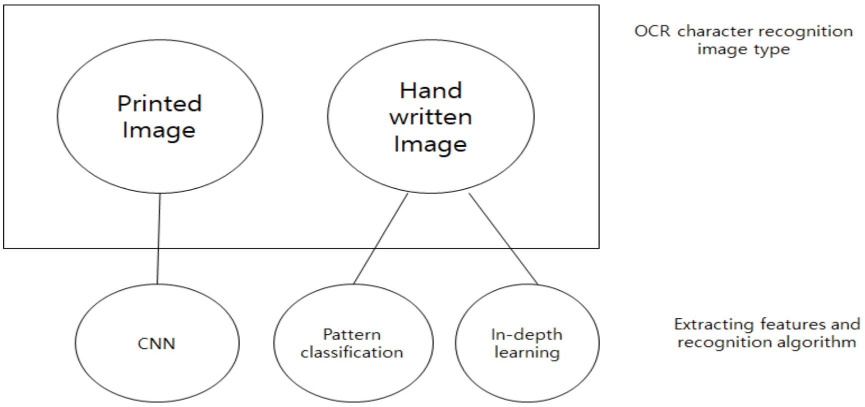
Printed Image
Hand written Image
OCR character recognition image type
CNN
Pattern classification
In-depth learning
Extracting features and recognition algorithm

Figure. 2. OCR Character Recognition Type and Algorithm

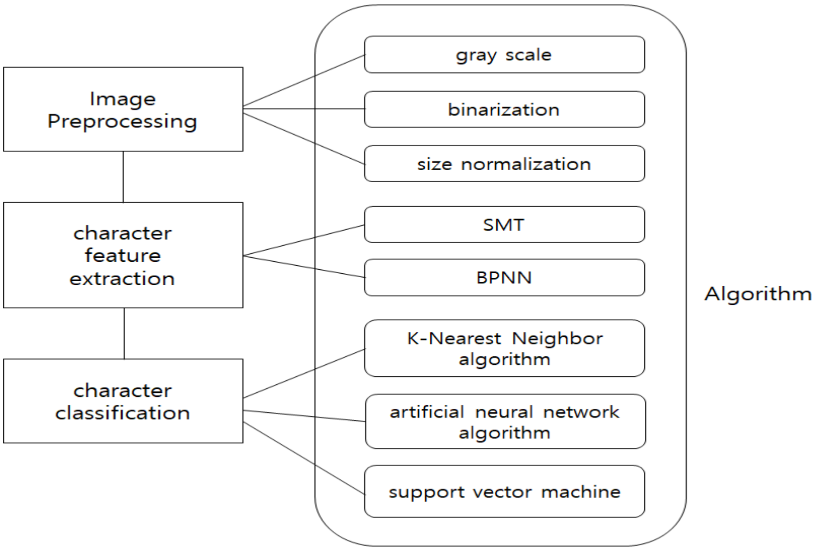

Figure. 3. Character Recognition Process and Algorithm

3. 2. Areas of Character Recognition

The using field of character recognition technology is expanding. Character recognition technology has been used in a wide range of fields for efficient processing and convenience of work in various fields. Table 1 shows the application of character recognition.

Table 1: Character Recognition Field

Application field	Explanation
Document work	▪ Character recognition for documents in the bank ▪ Character Recognition for Receipts ▪ Character Recognition for Passports ▪ Character Recognition for a specific field of a customer's document at an insurance company
License plate recognition	▪ Automatic license plate recognition for vehicles entering and leaving public institutions
Organize with text characters in printed documents	▪ Save text by extracting text characters from printed documents

4. OCR PROGRAM

Recently, due to the development of OCR-related programs, the number of popular programs is increasing. Popular OCR programs are such as Readiris, Abbyy Fine Reader, and Microsoft One Note. Table 2 below summarizes these OCR programs.

4. 1. Readiris

Readiris Pro is one of the most powerful OCR packages for the PC. In just a few seconds, you can turn your document into editable text. In fact, Readiris Pro contains more features than you need, but the most important point is that it is very accurate. Readiris Pro also provides an excellent way to maintain formatting. Readiris Pro faithfully reproduces the document's original format, replacing it with text, tables, and graphics columns in the output file. Except for some languages, this supports a wide range of languages (up to 130 languages). In particular, support for ASEAN languages continues to be made. Readiris can recognize all kinds of documents. You can specify the language you want to recognize. Another feature of the Readiris program is the ability to recognize all files in folders on your computer. The Readiris program can recognize any number of pages.

4.2. ABBYY FineReader

ABBYY FineReader provides powerful OCR, PDF viewing and editing for all types of PDF documents, including paper documents. ABBYY FineReader's OCR technology not only recognizes text quickly and accurately, but also preserves the original formatting of the document. ABBYY FineReader preserves the structure of the original document, including forms, hyperlinks, email addresses, headers, footers, captions, page numbers, and footnotes.

ABBYY FineReader's built-in text editor can compare the recognized text in the original image and change the content or format if necessary. You can manually specify the area of the image to capture and train the program to recognize special fonts that are not used often.

ABBYY FineReader supports 179 recognition languages. The program also intelligently detects the languages used in the documents. There is no need for change settings prior to the scanning.

4. 3. One Note OCR

One Note has an OCR function. It can be used as a technique for extracting text from pictures or images. One of the features of OneNote is that it originally provides OneNote OCR. Microsoft OneNote OCR is an OCR feature added by Microsoft to OneNote that allows users to recognize text in pictures, captures, and PDF prints. Simply you can select a picture or page, copy the text, and paste it into OneNote or another text processing tool.

When you scan the document using the utility it will be automatically OCR the scanned images and send the recognized text to your version of Word you've installed. To convert handwriting to text in OneNote, you first select the note that you want to convert. OneNote will then convert the handwriting in the note to typed text [6, 7].

4. 4. SimpleOCR

SimpleOCR is a popular freeware OCR software used by hundreds of thousands of users worldwide. SimpleOCR is also a royalty-free OCR SDK. If you have a scanner and don't want to retype the document, SimpleOCR is a fast and free way to do it. SimpleOCR freeware is free and can be used by anyone with no restrictions [8, 9, 10]. With **Simple OCR**, you could easily and accurately convert that paper document into editable electronic text.

4. 5. Tesseract

Tesseract is an open source OCR engine developed by HP. Tesseract was developed as a software and hardware add-on for HP's line. Tesseract has significantly improved accuracy compared to existing technology, but has not been commercialized. The next step for development is to study compression OCR at HP Labs Bristol. HP release Tesseract for open source.

Google has been sponsoring the project since 2006. As of 2018, it is evolving into a powerful OCR tool with built-in deep learning capabilities.

Table 2: OCR Program Feature

Products	Feature
READIRIS	▪ Very accurate ▪ reproduces the document's original format ▪ support for ASEAN languages ▪ recognize all files in folders on computer
ABBYY FineReader	▪ provides powerful OCR ▪ recognizes text quickly and accurately ▪ preserves the original formatting ▪ supports 179 recognition languages
One Note OCR	▪ extracting text from pictures or images ▪ work with Microsoft OneNote
SimpleOCR	▪ popular freeware OCR software ▪ SimpleOCR is a fast and free
Tesseract	▪ open source OCR engine developed by HP ▪ built-in deep learning capabilities ▪ greatly increase the recognition rate recent

5. CONCLUSION

This paper introduces the character recognition technology used as the core technology in the era of the 4th industrial revolution. Character recognition is used and utilized in many fields for convenient and fast data processing in our daily life. This paper introduces the character recognition technology. This paper introduces an overview and explanation of the basic concepts in character recognition. In the character recognition process, the binarization process, noise reduction, and region separation will be described. In addition, there are processes of separating characters from the image area, extracting features, extracting recognized characters, and finally, post-processing to improve accuracy. These processes are important technical elements in character recognition.

In addition, we introduce the programs used in the text recognition. This paper introduces various OCR programs that have been developed and used to recently. I introduces ABBYY FineReader, Readiris, one note OCR, simple OCR, Tesseract, etc.

REFERENCES

[1] Suhyun Kim, Songhee Lee, Sang Jun Lee, Sang Ho Lee, "Household storage service through Optical Character Recognition (OCR)", KOREA INFORMATION SCIENCE SOCIETY, pp. 377-379, 2017.12.

[2] Won-Yong LEE, "Development of character recognition system based on the image processing techniques", The Society of Convergence Knowledge Transactions, vol. 5, no. 2, pp. 99-103, 2017.7.

[3] Nameirakpam Dhanachandra, Khumanthem Manglem, Yambem Jina Chanu, "Image Segmentation using K-means Clustering Algorithm and Subtractive Clustering Algorithm", Eleventh International Multi-Conference on Information Processing-2015, pp. 764-771, vol. 54, 2015.

[4] Mehdi Rizvi, Hasnain Raza, Shahab Tahzeeb, "Optical Character Recognition Based Intelligent Database Management System for Examination Process Control", Proceedings of 2019 16th International Bhurban Conference on Applied Sciences & Technology (IBCAST), pp. 500-507, Jan., 2019.

[5] Zheheng Rao, Chunyan Zeng , Minghu Wu, Zhifeng Wang, Nan Zhao, Min Liu, Xiangkui Wan, "Research on a handwritten character recognition algorithm based on an extended nonlinear kernel residual network", KSII TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS VOL. 12, NO. 1, Jan. 2018.

[6] Nicole Dalia Cilia, Claudio De Stefano, Francesco Fontanella, Alessandra Scotto di Freca, "A ranking-based feature selection approach for handwritten character recognition", Pattern Recognition Letters, vol. 121, 15, pp. 77-86, Apr. 2019.

[7] Raymond Ptucha, Felipe Petroski Such, Suhas Pillai, Frank Brockler, Vatsala Singh, Paul Hutkowski, "Intelligent character recognition using fully convolutional neural networks", Pattern Recognition, vol. 88, pp. 604-613, Apr. 2019.

[8] Yael Fogel, Naomi Josman, Sara Rosenblum, "Functional abilities as reflected through temporal handwriting measures among adolescents with neuro-developmental disabilities", Pattern Recognition Letters, vol. 121, 15, pp. 13-18, Apr. 2019.

[9] Won-Yong LEE, "Development of character recognition system based on the image processing techniques", The Society of Convergence Knowledge Transactions, Vol. 5, No. 2, pp. 99-103, 2017.7

[10] https://www.simpleocr.com/