ISSN 2278-3091

Volume 9, No.3, May - June 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse47932020.pdf

https://doi.org/10.30534/ijatcse/2020/47932020

OCR Related Technology Methods

Bharti Sharma[1], Ashutosh Kumar Rao[2]

[1]Research Scholar, Sunder Deep Engineering College, Ghaziabad, India, bhartisharma0811@gmail.com

[2]Assistant Professor, Sunder Deep Engineering College, Ghaziabad, India, ashutoshrao7@gmail.com

ABSTRACT

The technology associated with character recognition has emerged as a vital technology within the era of the fourth historic period. Character recognition is developing as a core technology needed in various fields. Character recognition is performed by extracting characters from a picture and recognizing the extracted characters. Character recognition technology has been continuously developed.

Recently, together with the event of the fourth historic period, character recognition technology has been used as a core technology in many places. This paper introduces the technology associated with character recognition and therefore the program for character recognition.

Key words: OCR, OCR Program, OCR Technology, OCR Technology Trend.

1. INTRODUCTION

In the era of the 4th historic period, 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 a picture scanner to convert them into machine-readable characters. Character recognition technology has been continuously researched and developed within the field of OCR. Various sorts of OCR related programs are developed and used. Some OCR programs are hospitable the general public for free of charge, while others is 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 massive problem is that the popularity rate can vary greatly counting on the user's language. This is often often caused by the specificity of the language.

In recent years, OCR programs that solve these technical problems more actively have emerged. With the event of those technologies, the employment of OCR programs has recently become popular. Document analysis and string extraction from natural images are the foremost basic and important issues for understanding documents and pictures. While text recognition is already available in many commercial products, analyzing and recognizing complex documents and natural images isn't easy to resolve yet. Finding and analyzing characters in various environments and extracting text characters accurately has become a crucial technology.

Although various OCR-related studies are conducted, there's a scarcity of papers on technologies associated with character recognition. Therefore, this paper introduces the related technologies and functions required for OCR and summarize the characteristics of every technology. Additionally, we introduce OCR SW program.

2. 2. HISTORY OF OCR TECHNOLOGY DEVELOPMENT

The history of OCR technology development began in 1928. It’s a personality recognition method using pattern matching. It compares several standard pattern characters and input characters prepared before and selects the foremost just like the quality pattern character because the corresponding character. Around 1955, devices for recognizing printed numbers were invented within the UK and therefore the us. In the 1960s, various researches were conducted around IBM within the us. The research on the advance of the character recognition rate, the handwritten recognition problem, etc. has been conducted. With in the third generation OCR system, which appeared within the mid-1970s, the standard of documents for character recognition was poor. The environment for OCR was primarily used as a tool for straightforward typing before the appearance of PCs.

As the element performance improved in 1980, software development for OCR progressed, and commercial systems began to spread. Because of the proliferation of computers within the 1980s, the necessities for character recognition began to extend. At now, they actively conducted researches on character recognition within the culture of Chinese characters. Recently, because of the spread of smart phones, research on the sphere of character recognition using smart phones has been actively conducted. The study of technology associated with character recognition is conducted mainly in West Germanic. This is often because English-related documents are main stream in our life. Recently, they expanded researches to review the character recognition of assorted languages round the world [1, 2, 3].

3.OCR RELATED TECHNOLOGY AND APPLICATION FIELD

Techniques associated with character recognition are required in various fields. Particularly, character recognition technology using image processing technology is rapidly developing. While these technologies are developing rapidly, there are many technologies that require to be addressed in computer recognition.

The process of character recognition is mostly divided into three stages: preprocessing, region analysis, and recognition. The method of preprocessing for character recognition is a crucial part to extend the character recognition rate. Within the program for character recognition, character recognition is performed using data that has been preprocessed. Character recognition also differs counting on language. This paper focuses on print English character, which is usually 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's recognizable through preprocessing. During this process, unnecessary information is removed and conversion for binarization is performed. Particularly, the handwriting character is usually informal, and thus, a tilt correction process is important [2]. A typical character recognition process is shown in Figure 1 below.

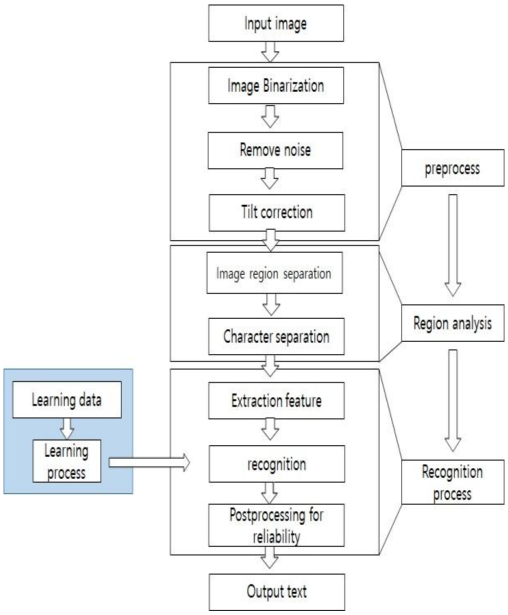

Figure 1: Character Recognition Process

Character reading will be divided into two parts: preprocessing of computer file and character recognition supported processed data. Preprocessing procedure is as in Figure 1.Image data goes through the binarization module, thinning module and line extraction module. Image data are processed and have an identical format, and so the information is suitable for character recognition through the gradient difference module, the thinned line separation module, the Bezier curve calculation module, and therefore the pen thickness calculation module. Then, the characters are recognized through the character attribute graph generation module, the essential stroke graph matching module, and therefore the character graph matching module [2].

Before processing the binarization process, we first undergo 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 rationale for binarization is to differentiate the background from the characters clearly. In binarization, the characters will be extracted accurately in keeping with the brink setting for distinguishing the characters from the background. There are various ways to search out the brink.Image segmentation is one in all the ways commonly accustomed classify the pixels of the image correctly within the video image. This divides the image into several distinct regions so the similarity of pixels in each region is high and therefore the contrast ratio between regions is high. Image segmentation is one in all the essential problems of image analysis. The importance and usefulness of image segmentation allows us to research images accurately using extensive research and several.

Other proposed approaches like intensity, color, and texture. There are various images - Segmentation techniques supported threshold, boundary, cluster, and neural network [2, 3, 4]. A very important a part of preprocessing for character recognition is removing noise. Noise is an unwanted pixel deformation of a picture that reduces the effectiveness of the image processing mechanism. Gaussian Blurring Smoothing technology is employed to get rid of small amounts of noise.

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

Through the method 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 especially used for handwritten character recognition. Feature extraction is processed supported original image information like color, structure, texture, projection histogram, and instantaneous invariant. Once feature extraction is complete, the classifier is employed 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 may be a important technology which will be applied to varied fields Within the 4th age because it enables text recognition from a given image. OCR technology even makes it possible to spot characters in handwritten scripts [4].

4.CHARACTER RECOGNITION

Character recognition process is as in Figure 1, Figure 4. Handwritten character recognition features a great potential for application, and there's a good demand in accordance with the event of society in industries like a picture recognition system or a handwriting data input device. Many researchers have an interest within the study of handwritten character recognition within the industry. Especially, within the field of image processing and pattern classification, handwritten character recognition has been extensively researched and developed. The methods currently used for handwriting recognition represent two categories: handwritten character recognition supported pattern classification and in-depth learning as in Figure 2.Character recognition may be a process of extracting features of varied images. One in every of the important technologies, Surface Mount Technology (SMT), or Back Propagation Neural Network (BPNN) is employed to handle character recognition [6].

Character recognition can find solutions through every pixel. This typically divides the image into regions that are homogeneous in some sense or centered on regions of significance. Other algorithms for segmentation like this include K-Means clustering, expectation maximization algorithm, and mean shift algorithm [4, 5].

Through the method of session development, it extracts the string from the image. It extracts the characters that correspond to the string and converts them into characters.

Approximate optimal solutions like heuristics algorithms thanks to various difficult problems. In general, the execution time is commonly a mathematical function within the character recognition algorithm. Therefore, they need conducted researches on character recognition methods using various methods. Some cases used Genetic Algorithms as probabilistic approach. Optical Character Recognition is principally focused on text recognition for printed documents.

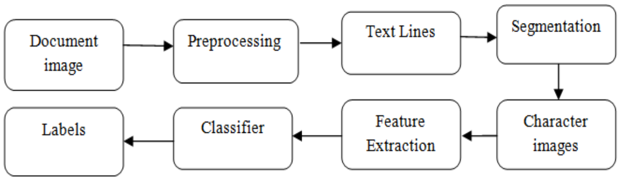

Figure 2: Flow Diagram of OCR

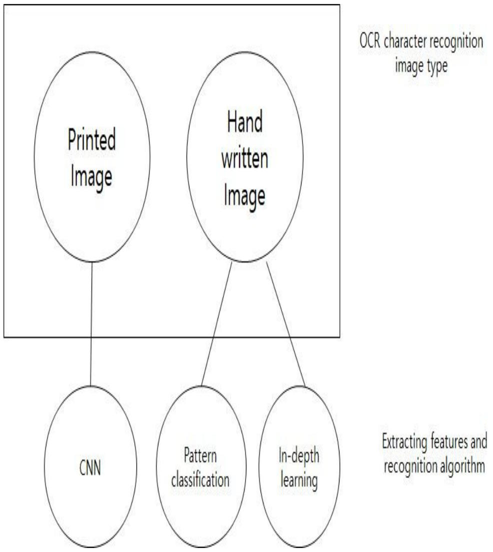
OCR character recognition image type
Printed Image
Hand written Image
CNN
Pattern classification
In-depth learning
Extracting features and recognition algorithm

Figure 3: OCR Character Recognition Type and Algorithm

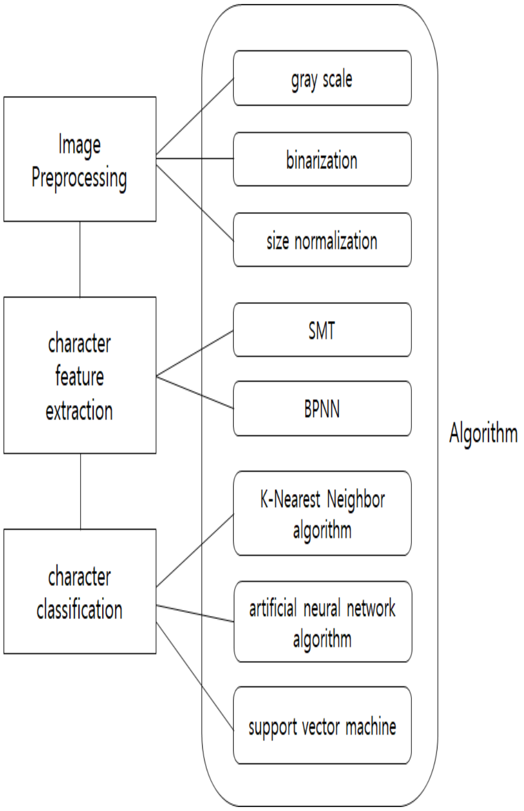

Figure 4: Character Recognition Process and Algorithm

5. AREAS OF CHARACTER RECOGNITION

The using field of character recognition technology is expanding. Character recognition technology has been employed in a good range of fields for efficient processing and convenience of labor in various fields. Table 1 shows the applying 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

6.OCR PROGRAM

Recently, because of the event of OCR-related programs, the quantity of popular programs is increasing. Popular OCR programs are like Readiris, Abbyy Fine Reader, and Microsoft One Note. Table 2 below summarizes these OCR programs.

6.1 Readiris

Readiris Pro is one amongst the foremost powerful OCR packages for the PC. in only a couple of seconds, you'll turn your document into editable text. In fact, Readiris Pro contains more features than you would like, but the foremost important point is that it's very accurate.

Readiris Pro also provides wonderful thanks to maintain formatting. Readiris Pro faithfully reproduces the document's original format, replacing it with text, tables, and graphics columns within the computer file. Aside from some languages, this supports a good range of languages (up to 130 languages). Above all, support for ASEAN languages continues to be made. Readiris can recognize all types of documents. You'll specify the language you would like to acknowledge. Another feature of the Readiris program is that the ability to acknowledge all files in folders on your computer. The Readiris program can recognize any number of pages.

6.2 ABBYY Fine Reader

ABBYY Fine Reader provides powerful OCR, PDF viewing and editing for all sorts of PDF documents, including paper documents. ABBYY FineReader's OCR technology not only recognizes text quickly and accurately, but also preserves the initial formatting of the document. ABBYY Fine Reader preserves the structure of the initial 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 within the original image and alter the content or format if necessary. You'll be able to manually specify the realm of the image to capture and train the program to acknowledge special fonts that aren't used often.

ABBYY Fine Reader supports 179 recognition languages. The program also intelligently detects the languages utilized in the documents. There's no need for change settings before the scanning.

4.3. One Note OCR

One Note has an OCR function. It will be used as a way for extracting text from pictures or images. One in all the features of OneNote is that it originally provides OneNote OCR. Microsoft OneNote OCR is an OCR feature added by Microsoft to OneNote that enables use to acknowledge text in pictures, captures, and PDF prints. Simply you'll select an image or page, copy the text, and paste it into OneNote or another text processing tool. When you scan the document using the utility it'll 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 initially select the note that you just want to convert. OneNote will then convert the handwriting within the note to typed text [6,7].

6. 4. Simple OCR

Simple OCR could be popular freeware optical character recognition software employed by many thousands of users worldwide. Simple OCR is additionally a royalty-free OCR SDK. If you have got a scanner and do not want to retype the document, Simple OCR could be fast and free thanks to have it off. Simple OCR freeware is free and may be employed by anyone with no restrictions [8, 9, and 10]. With Simple OCR, you'll easily and accurately convert that paper document into editable textual matter.

6.5 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. Successive step for development is to review compression OCR at HP Labs Bristol. HP release Tesseract for open source.

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

7. CONCLUSION

This paper introduces the character recognition technology used because the core technology within the era of the 4th age. Character recognition is employed and utilized in many fields for convenient and fast processing in our lifestyle. This paper introduces the character recognition technology. This paper introduces a summary and explanation of the fundamental Concepts in character recognition. With in the character recognition process, the binarization process, noise reduction, and region separation are described. Additionally, there are processes of separating characters from the image area, extracting features, extracting recognized characters, and eventually, post-processing to enhance accuracy. These processes are important technical elements in character recognition.

In addition, we introduce the programs employed in the text recognition. This paper introduces various OCR programs that are developed and wont to recently. I introduce ABBYY Fine Reader, Readiris, one note OCR, simple OCR, Tesseract, etc.

ACKNOWLEDGEMENT

It’s a pleasure and a great blessing of GOD for working on the project named “OCR Related Technology Methods”. Where in we gained knowledge by working under the able leadership of our **Head of department Mr. Ashutosh Kumar Rao** who helped and supported us in every sphere of our project. We all thank our **Project in- charge Mr. Ashutosh Kumar Rao** who well supported us and provided us with his precious time and support Besides that we thank whole of the CSE department for their appreciation and kind support.

REFERENCES

1. S. Kim, S. Lee, S. Jun Lee, S. Ho Lee, **“Household storage service through Optical Character Recognition (OCR)”**, *korea information science society*, pp. 377-379, 2017.12.
2. W.Y 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. N. Dhanachandra, K. Manglem, Y. J. Chanu, **“Image Segmentationusing K-means Clustering Algorithm and Subtractive Clustering Algorithm”,** *Eleventh International Multi-Conference on Information Processing*-2015, pp. 764-771, vol. 54, 2015. https://doi.org/10.1016/j.procs.2015.06.090
4. M. Rizvi, H. 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. https://doi.org/10.1109/IBCAST.2019.8667127
5. Z. Rao, C. Zeng , M. Wu, Z. Wang, N. Zhao, M. Liu 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. https://doi.org/10.3837/tiis.2018.01.020
6. N.D. Cilia, C.D Stefano, F. Fontanella, A. Scotto di Freca, **“A ranking-based feature selection approach for handwritten character recognition”**, *Pattern Recognition Letters*, vol. 121, 15, pp. 77-86, Apr. 2019. https://doi.org/10.1016/j.patrec.2018.04.007
7. R. Ptucha, F. PSuch, S.Pillai, F. Brockler, V. Singh, P. Hutkowski, **“Intelligent character recognition using fully convolutional neural networks”**, *Pattern Recognition*, vol. 88, pp. 604-613, Apr. 2019. https://doi.org/10.1016/j.patcog.2018.12.017
8. Y. Fogel, N. Josman, S. 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. https://doi.org/10.1016/j.patrec.2018.07.006
9. W.Y. LEE, **“Development of character recognition system based on the image processing techniques”**, *The Society of Convergence Knowledge Transactions*, Vol. 5, pp. 99-103, 2017.7
10. Https://www.simpleocr.com/