Optical character recognition with different languages

Yifan Wu[1,†] and Yuxi Zhang[2,3,†]

[1]Department of Scotland, University of Wuxi Taihu, Rongxiang Street, Wuxi 214063, China
[2]Department of Mathematics, University of Manchester, Oxford Road, Manchester M13 9GL, UK

[3]631402070312@mails.cqjtu.edu.cn
[†]These authors contributed equally.

Abstract. Optical character recognition is the combination of optical technology and computer technology to identify text in an image and then recognize the text content in the image, providing individuals with a great deal of ease in their daily lives. Document text recognition, natural scene text recognition, bill text recognition, and ID card recognition have been used in daily life, but there are still many factors that lead to inaccurate identification and detection. Therefore, different texts, patterns or characters are suitable for different types of Optical character recognition. In this paper, we can learn about the Optical character recognition operation methods and find the similarities and differences through researching the technical routes and four different types of Optical character recognition. In addition, by comparing the Optical character recognition of several commonly used languages, the advantages and disadvantages of each method can be analysed.

Keywords: optical character recognition, different languages, advanced technology.

1. Introduction

Optical character recognition (OCR) refers to the process of analysing and identifying image files of text materials to obtain text and layout information. The main development history of OCR can be seen in figure 1. OCR was first proposed by German scientist Tausheck in 1929, but it was not implemented until the computer was invented in 1946. IBM (International Business Machines Corporation) was the first to study printed Chinese character recognition, and in 1966 they published the first article on Chinese handwritten character recognition. In the early stage, the numbers were used as the object of research to identify the postcodes on the mail, and it was used to deliver mail by region. So far, the postcode has always been the address writing method advocated by various countries. Nowadays, printing, handwriting, numbers, symbols have all been studied. So optical character recognition has been widely used in every aspect of life. The main indicators to measure the performance of an OCR system are rejection rate, false recognition rate, recognition speed, user interface friendliness, product stability, usability and feasibility, etc.

According to the recognition scene, OCR can be roughly divided into a dedicated OCR for recognizing a specific scene and a general OCR for recognizing a variety of scenes. The techniques of using dedicated OCR are mature now. But there are still some problems remaining when implementing general OCR. The reason why natural scene text recognition is extremely difficult is that the background of image is rich and varied, which leads to distorted text layout and inconsistent font styles, various colours, and arbitrary directions, even exposure, reflections, and partial occlusion can cause problems for recognizing images.

The practical applications of OCR are seen in document text recognition, which can help electronically manage paper books, newspapers, magazines and historical literature to preserve the documents. In healthcare, patient records can be processed and updated in real time. Moreover, bill text recognition can avoid financial staff from manually entering a large amount of bill information. For bank staff, text recognition can process and verify documents for financial transactions, which can improve transaction security. The OCR can also be used in identification recognition to save the time and cost of checking people's identities. ID card recognition can quickly identify identity information such as ID card, bank card, driver's license, etc., and directly convert the text information of the document into editable text, which is convenient for real-time identity verification of relevant personnel. Many things are simplified because of implementing OCR and this can help the society work more efficiently. We have a bird's-eye view of the history of OCR development and summarize the application of OCR in different language recognition. Figures and tables summarize the development history and application process. In the methods section, different methods are compared, and current trends are drawn. The difficulty of this research is to study and understand each method.

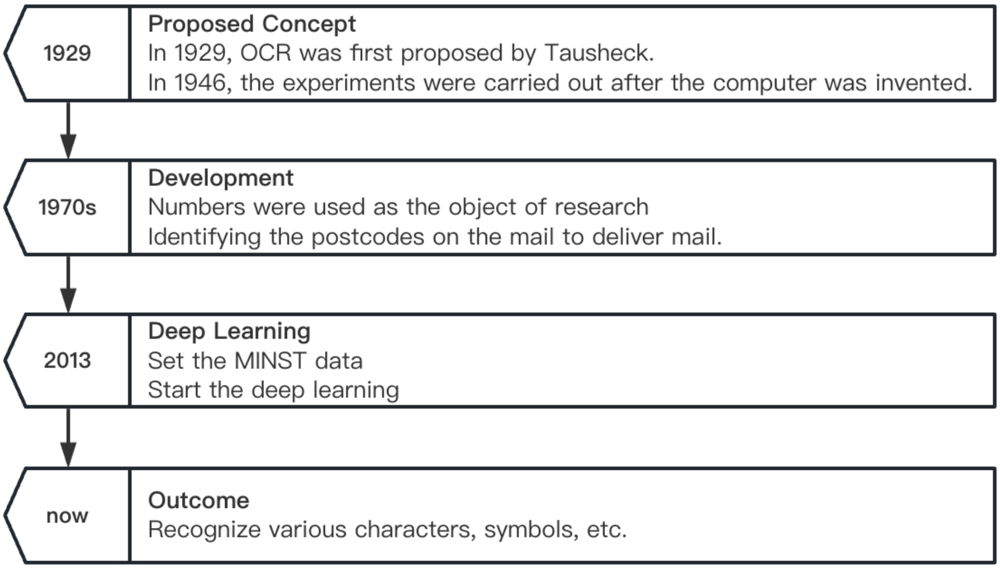
1929
Proposed Concept
In 1929, OCR was first proposed by Tausheck.
In 1946, the experiments were carried out after the computer was invented.
1970s
Development
Numbers were used as the object of research
Identifying the postcodes on the mail to deliver mail.
2013
Deep Learning
Set the MINST data
Start the deep learning
now
Outcome
Recognize various characters, symbols, etc.

Figure 1. The development of the optical character recognition.

2. Method

The OCR operation process includes only five steps and is not complicated, but each part has a high burden. Although many versions of OCR work in different ways, there are still certain commonalities. Moreover, the discipline of OCR focuses on four different languages: English, Chinese, South Indian, Tamil and Arabic.

2.1. OCR technology route

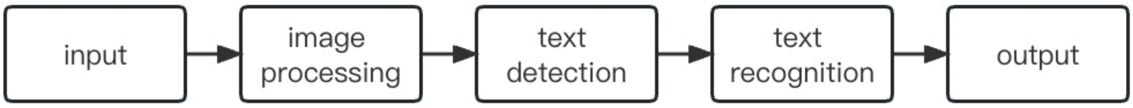

Figure 2. The route of OCR.

The OCR process can be seen from figure 2. It is divided into five steps: input, image pre-processing, text detection, text recognition and output. First of all, the input section reads and transforms the articles that need to be recognized into binary data while the OCR programme analyses the scanned images. The bright area is the back ground, while the dark one is the words. Furthermore, pre-processing is where the OCR software cleans up the picture and eliminates flaws in preparation. There are several cleaning techniques such as applying a slight offset correction to the scanned document to fix alignment issues or cleaning up borders and lines in the image or removing image speckle. Additionally, text recognition is mainly responsible for pattern matching and feature extraction. Pattern matching separates character images (called glyphs) and compares them to similar stored glyphs. Feature extraction segments or decomposes glyphs into features such as line, closed loop, line direction, and line focus. Pattern matching is more suitable for scanned images of documents with stored font input, while feature extraction is suitable for finding the best match or the closest matching glyph. Eventually, after finishing analysing, the system turns the acquired data into digital files during post-processing. For example, annotated PDF files which can be created by some OCR systems contains before and after scan versions of scanned documents.

2.2. Types of OCR

Simple OCR software: The engine stores several distinct fonts and text image patterns as templates. OCR software compares an image of words with its internal database using a pattern-matching algorithm, character by character. If the system matches the text verbatim, it is referred to as optical text recognition. This solution is limited because there are so many fonts and handwriting styles that each specific one cannot be recorded and saved in the database.

Intelligent Character Recognition Software (ICR): This technology can read text like a human. Machine learning systems, also known as neural networks, analyse text at multiple levels and repeatedly process images. It searches for various image properties (such as curves, lines, intersections, and loops) and integrates the findings of all these several layers of analysis to provide final results. Despite the fact that ICR can normally process only one scanned character at one time, results are available in seconds, which is fast.

Intelligent Text Recognition: works the same as ICR, but it processes the entire text image instead of pre-processing the image into characters.

Optical Mark Recognition: recognizes logos, watermarks and other text symbols in documents.

2.3. Different languages

2.3.1. English. English is the world's most frequently spoken language. It is the universal language of many nations, also bilinguals utilize it as a second language. Neural Networks or Harmonic Markov Models may integrate statistical and structural information for numerous character patterns. Deep neural networks have recently become popular. Convolutional Neural Network architecture is a deep neural network architecture that outperforms most advanced visual stimulus or input categorization results. In a recent study, Fully Convolutional Neural Network were utilised to obtain Character Error Rates and Word Error Rates of 4.7%, 8.22%, 2.46%, and 5.68% separately [1]. Jayasundara suggested a unique approach for handwritten character identification with extremely tiny datasets termed Capsule Networks [2]. According to the research, when applied to tiny datasets, the proposed technique obtains an accuracy rate of 90.46%.

2.3.2. Chinese. In the year 2000, one of the earliest Mandarin language research projects was accomplished. The researchers created an adaptive user module for character identification and personal adaptation using Probabilistic Neural Networks. In 10 adaptation cycles, the resultant recognition accuracy reached 90.2%. HCL2000 proposed in 2009 is a handwritten database of Chinese Character which contains close to 4000 commonly used characters as well as the information of the individual authors [3]. Three separate methods were used to evaluate HCL2000: Linear Discriminant Analysis, Locality Preserving Projection and Marginal Fisher Analysis [4]. Before analysing, the Nearest Neighbour classifier allocates the input picture to a group of characters. The outcomes demonstrate that Marginal Fisher Analysis and Locality Preserving Projection are better than Linear Discriminant Analysis.

In 2013, categorization on extracted feature data, the researchers explored whether all the online and offline isolated character and handwritten text could be recognized. Techniques include Convolutional Neural Network, Linear Discriminant Analysis, Modified Quadratic Discriminant function, Compound Mahalanobis Function and Multilayer Perceptron in the experiment [5]. Research shows that neural network-based methods were superior at identifying both isolated character and manuscript text. Five years later, Chinese script researchers employed neural networks to detect CAPTCHA, medical document recognition, licence plate identification, and historical documents [5].

2.3.3. South Indian. Malayalam is a type of South India, and its structure is complex. Back Propagation Neural Networks is used to make recognition in common language. But in Malayalam there are many combinational letters that BPNN cannot handle. Therefore, the method used for this language to make separation is labelling the image [6]. Labelling is a method that uses the four and eight connected pixels of an image. Another OCR method of recognizing Malayalam is proposed through the CDAC [6]. This technique uses Otsu's algorithm for binarization. It is reported that this method has an accuracy of 97%.

2.3.4. Tamil. Tamil is a language which has strange structure so that it is needed to use classification while making recognition. Support Vector Machine is a classier method that constructs hyperplanes in a multidimensional space to display classification [7]. Different letters are separated in the multidimensional space. SVM can handle multiple continuous and categorical variables because it supports regression and classification. SVM uses an iterative training algorithm. This algorithm can construct a hyperplane since it can find the minimize in the loss function.

2.3.5. Arabic. Throughout the last few decades, the development of handwritten Arabic OCR systems has taken various steps. In the beginning of 21st century, research concentrated on neural network approaches for recognition and the creation of database variations [8]. Pechwitz created the first IFN/ENIT-database in 2002 to mock test Arabic OCR systems [9]. With over 470 citations, this is one of the most referenced datasets. In 2009, Graves and Schmidhuber invented a worldwide offline handwriting recognizer accepting raw pixel data as an input on the basis of multidirectional recurrent neural networks and connectionist temporal classification. The system has a total accuracy of 91.4% [10].

In 2018, the DCNN (deep CNN) technique was used to detect the handwritten Arabic characters in the case of offline. The DCNN technique employing transfer learning was 98.86% accurate for both datasets [11]. Moreover, Histograms of Oriented Gradient for trait abstraction and Support Vector Machine for classifying characters were applied on the manuscript data set during another experiment. This collection comprises 99% accurate city, town, and village names [12].

3. Conclusion

In this paper, the OCR technology and its development are studied in detail. We summarize the applications of OCR in several language usage situations, and understand some new technologies based on OCR. We tend to find the best one, but it is difficult because each language has its own idiosyncrasies and writing styles. OCR still has a lot to be improved, therefore, we hope that in the future there will be an OCR technology that can be applied to all languages.

References

[1] Ptucha, R., Such, F. P., Pillai, S., Brockler, F., Singh, V., & Hutkowski, P. (2019). Intelligent character recognition using fully convolutional neural networks. *Pattern recognition*, 88, 604-613.

[2] V. Jayasundara, S. Jayasekara, H. Jayasekara, J. Rajasegaran, S. Seneviratne and R. Rodrigo, 2019 "TextCaps: Handwritten Character Recognition with Very Small Datasets," *IEEE Winter Conference on Applications of Computer Vision (WACV),* Waikoloa, HI, USA, pp. 254-262, doi: 10.1109/WACV.2019.00033.

[3] H. Zhang, J. Guo, G. Chen and C. Li, 2009 "HCL2000 - A Large-scale Handwritten Chinese Character Database for Handwritten Character Recognition," *2009 10th International Conference on Document Analysis and Recognition*, Barcelona, Spain, pp. 286-290, doi: 10.1109/ICDAR.2009.15.

[4] Huang Libo, Ling Yongquan. 2021 A parameter-free local linear discriminant analysis method[J]. *Computer Science and Applications*, 11(4): 1042-1052. https://doi.org/10.12677/CSA.2021.114107

[5] J. Memon, M. Sami, R. A. Khan and M. Uddin, "Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)," *in IEEE Access, vol. 8, pp. 142642142668, 2020,* doi: 10.1109/ACCESS.2020.3012542.

[6] Rahiman, M.A. and Rajasree, M.S. (2009) “A detailed study and analysis of OCR Research in South Indian scripts,” *2009 International Conference on Advances in Recent Technologies in Communication and Computing [Preprint]*. Available at: https://doi.org/10.1109/artcom.2009.45.

[7] Seethalakshmi, R. *et al.* (2005) “Optical character recognition for printed Tamil text using Unicode,” *Journal of Zhejiang University-SCIENCE A*, 6(11), pp. 1297–1305. Available at: https://doi.org/10.1631/jzus.2005.a1297.

[8] N. Mezghani, A. Mitiche and M. Cheriet, 2002, "On-line recognition of handwritten Arabic characters using a Kohonen neural network," *Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition,* Niagra-on-the-Lake, ON, Canada, pp. 490-495, doi: 10.1109/IWFHR.2002.1030958.

[9] Pechwitz, M., Maddouri, S. S., Märgner, V., Ellouze, N., & Amiri, H. (2002, October). IFN/ENIT-database of handwritten Arabic words. In *Proc. of CIFED* (**Vol. 2**, pp. 127-136). Citeseer.

[10] Graves A, Schmidhuber J. 2008 Offline handwriting recognition with multidimensional recurrent neural networks[J]. Advances in neural information processing systems, 21.

[11] C. Boufenar and M. Batouche, 2017 "Investigation on deep learning for off-line handwritten Arabic Character Recognition using Theano research platform," *2017 Intelligent Systems and Computer Vision (ISCV)*, Fez, Morocco, pp. 1-6, doi: 10.1109/ISACV.2017.8054902.

[12] R. A. Khan, A. Meyer, H. Konik and S. Bouakaz, "Pain detection through shape and appearance features," *2013 IEEE International Conference on Multimedia and Expo (ICME)*, San Jose, CA, USA, pp. 1-6, doi: 10.1109/ICME.2013.6607608.