A Study of The OCR Development History and Directions of Development

Junmiao Wang

Nanjing University of Posts and Telecommunications Nanjing, China

b19050508@njupt.edu.cn

Abstract. Optical character recognition (OCR) is a well-established technology that enables the conversion of scanned images and documents into editable and searchable electronic text. This technology has numerous applications across a range of industries and has proven to be a crucial tool for digitizing books, documents, and records. One of the main benefits of OCR technology is its ability to automate data entry processes, saving time and reducing errors that can be introduced through manual data entry. Additionally, OCR technology is used to extract text from images, which can be used as input data for machine learning and artificial intelligence applications. To better understand the current state of OCR technology, this systematic literature review is conducted that collected various research articles on the topic. The results of this paper provide insights into the strengths and limitations of OCR technology and also offer directions for future research in the field. In terms of language support, the paper found that OCR technology is capable of supporting a wide range of languages, including English, Spanish, Chinese, and many others. However, the accuracy of OCR technology can vary greatly based on the language, with some languages being more challenging to recognize than others. With continued advancements in technology and the increasing need for digitization, OCR technology will continue to play a crucial role in the development of many industries.

Keywords: optical character recognition, deep learning, text extraction.

1. Introduction

Optical character recognition (OCR) allows computers to read and process text which appears in handwriting images or some scanned documents. To analyze the patterns of the characters in an image and converts them into machine-readable text, making it feasible to edit and recognize. Today, this technology has a wide range of applications, including digitizing books and documents, extracting texts from images, etc.

When this technology comes out initially in the 1950s, scanner was used to transfer books or documents to images, using early OCR to read some regular texts [1]. But early OCR systems were limited in their accuracy as well as only a small set of predetermined characters can be recognized.

With the development of more sophisticated algorithms and computer vision techniques, OCR began to fix the problem that happened in the early system. In addition, the number of characters can be recognized got a massive increase. Meanwhile, researchers try to make the OCR system can adapt different fonts, not just be limited to one or two.

With the gradual spread of artificial intelligence, technology is used in many fields. OCR technology is closely related to machine learning and artificial intelligence, as it involves the use of algorithms to analyze and recognize patterns in data. Some new OCR systems with machine learning expand the applications, i.e. the use of large volumes of text data for training machine learning models and natural language processing. The latest generation of OCR technology has greatly improved in terms of accuracy and versatility compared to earlier versions. Modern OCR is able to transcribe a wide range of different languages and fonts, including non-Latin text. As various neural networks and algorithms which include Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, recurrent Neural Networks (RNN), etc. are applied to OCR, this technology increases the recognition accuracy. some of the more mature neural network models such as CNN and RNN have successfully made the handwriting recognition rate of OCR improve again and can even reach 99% or higher. Neural network-based OCR has also been gradually developed to

recognize handwritten characters in more complex languages, and the recognition speed has also increased dramatically.

This paper also analyzes the accuracy of OCR technology, finding that the technology has greatly improved over the years and is capable of achieving high levels of accuracy for certain types of documents and images. However, there are still many challenges to overcome, including recognizing text in low-quality images and dealing with handwriting recognition. This paper looked at the models used for OCR technology, finding that a variety of models have been developed and used, including neural network models, decision tree models, and more.

This paper gives a brief overview of the development process of OCR and the remarkable methods in Section 2 and describes the classification of OCR in Section 3, which is divided into language classification, dataset classification, and method classification. finally, this paper draws conclusions in Section 4.

2. Method

Since the concept of OCR technology emerged, different researchers have proposed different methods, each of which has its feature.

Table 1 below shows a handful of researchers' methods they used, analyzing the features and disadvantages.

Table. 1 Method used in research.

Year	Method	Feature	Drawback
1956	Convert the two-dimensional information [2]	Small calculation volume	Limited recognition of character
1962	electronics and optical techniques [3]	Recognize all characters of English	No handwriting recognition
1957	Combination with peephole method [4]	High speed of recognition	
1961	Combination with vertical scan [5]	Different Scan method	
1968	Template Matching and Structural Analysis [6]	Reduction of binary patterns calculation	
1982	Standardization and automatic recognition [7]	Hand-writing recognition	Limited recognition rate
1988	Combination with Bayesian classifier [17]	High accuracy	
1990	complementary algorithms [18]	Reinforce the recognition	No handwriting recognition
1998	graphical method [10]	Classified using graph similarity measure	Not mature
2019	Classical Grammar based [11]	Find the similarity in graphs	No grammar restriction
2005	8-directional features [19]	Chinese character recognition	Not mature
2009	Linear Discriminant Analysis (LDA) [12]	large-scale handwritten Chinese Character	Not mature
2018	Convolutional Neural Network (CNN) [13]	recognize CAPTCHA	Low recognition accuracy
2019	BIO tagging [20]	information propagation	Not mature
2022	Generative Adversarial Network (GAN) model [14]	Recognize degraded document images	Poor segmentation results

In the early stage of research, the technology could only recognize a little standard font, and the recognizable languages were also limited, most research select English, because there were fewer characters need to be recognized in English than in other languages. Different from now, OCR technology used before 2000 more inclined to segment characters, or perform dimension reduction transformation to reduce the difficulty of computer analysis and find out the feature points of different characters in the character set, using this feature to recognize characters.

The direction of OCR development has changed a lot. And the functions are getting increasingly comprehensive. Kelner and Glauberman use a slit to scan characters from top to bottom, converting the image to one-dimension, so that researchers only need to do some simple calculations to acquire the area of the black portion [2]. In fact, this method provides a more primitive idea, but this method is extremely limited to the character set, recognizing only a small number of characters and only specific fonts. Hannan and Solatron Electronics Group Ltd were doing the same thing with some dissimilar methods [3].

The n, Weeks et al. used a method called the structure analysis method which has a little different from the method that Kelner et al. used [5]. Researchers used this method to recognize some stylized fonts, and some of them were applied to try to recognize handwriting characters. Sakai et al. also create a model using an asynchronous reading method based on structure analysis [6]. This research greatly improves the accuracy of recognition.

Although some research tried to find a method to recognize unformal font characters, like handwriting characters, most researchers are inclined to improve the accuracy of formal fonts. Suen et al. are one of the researchers who develop the OCR technology on the recognition of handwriting. Although the recognition rate is limited [7], researchers realized that the recognition of handwritten fonts is also a necessary direction for the development of OCR technology from this time on.

To further enhance the OCR, researchers create many approaches to recognize characters in the past forty years. After that, a concept of neural network appeared, Krzyzak et al. made a model composed of several layers of interconnected elements, trying to utilize the model to generate an OCR which is consisted of the neural network [8]. Actually, researchers have already studied this direction for a long time. OCR with neural networks is not perfect initially, but with more and more features being developed, this technology has already matured. Hull et al. and Nadal et al. also participate in the development of neural network OCR and propose some new features respectively [9].

From that on, researchers have focused on neural networks, creating different machine learning methods, such as Support Vector Machine (SVM), k Nearest Neighbor (kNN), Decision Tree (DT), etc. With the development of Computer calculation speed, some deep learning models are presented, such as Convolutional Neural Network (CNN), etc. S. Lavirott et al. are all using structural pattern recognition method, classifying objects based on pattern structures [10]. S. Lavirott et al. employ a distinct model, utilizing the graphical method which consists of pairs of nodes (N) and edges (E).A. Chaudhuri used strings and trees to represent models based on grammar, improving the OCR recognition rate since analyzing the context of the article.

After researchers start using neural networks, some of them try to use this technology to recognize different characters. Zhang et al. have developed an OCR system incorporating Linear Discriminant Analysis (LDA), Locality Preserving Projection (LPP), and Marginal Fisher Analysis (MFA), capable of recognizing large-scale databases of handwritten Chinese characters [11]. Meanwhile, D. Lin al. attempt to use CNN to reinforce the system, this eventually allows the system to try to recognize some CAPTCHAs, although the frequency of recognition is not high and usually depends on the complexity of the image [12].

In recent years, OCR technology has gradually matured, so some researchers began to shift their research direction to using OCR technology to identify the text of fragmented documents. As in the work done by Ayan Chaudhury, they use OCR technology, paired with a neural network system that finally recognizes degraded document images [13]. This is certainly a suitable research direction for the future development of OCR.

3. Classifacation

The development of OCR technology has gone through decades, so researchers are bound to produce different research directions, and this section elaborates on the relevant classification of OCR technology.

3.1. Languages

Researchers come from different countries try to exert this technology in different languages. A survey of articles in literature databases, organized by time and language type, shows that the development of OCR is improving year by year. Most of the initial studies have focused on English as the main recognition language, but in recent years the development has gradually favored languages with multiple types of characters, such as Chinese, Indian, Arabic, etc. The United Nations Educational, Scientific and Cultural Organization (UNESCO) has reported that a minimum of 43% of the languages spoken worldwide are at risk of extinction, according to the report entitled "World's Languages in Danger." [14]. Using OCR technology, along with NLP natural language recognition systems, is a viable and effective way to protect these languages at risk.

Among the most widespread use of the system in English, many sophisticated methods have been used in business, and recognition speed and accuracy have met the needs of the market.

3.2. Dataset

Standardized datasets are typically used to evaluate and compare various OCR algorithms, in order to make meaningful comparisons. he main focus of optical character recognition research is on six languages: English, Arabic, Indian, Chinese, Urdu, and Persian/Farsi. Therefore, there are publicly available datasets for these languages, such as CEDAR, MNIST, HCL2000, etc.

Most of these character sets are handwritten character sets. Because, for the recognition of standard fonts, the OCR technology in recent years has been relatively mature, thus the standard character sets are no more than necessary.

There are numerous datasets used to study OCR techniques, just some of the more commonly used datasets are listed here.

3.2.1 CEDAR

The University of Buffalo developed a dataset recognized as the first substantial collection of handwritten characters. The images in CEDAR are scanned at a high resolution of 300 dots per inch (dpi), ensuring sharp and clear representation of the handwritten characters. This high scanning resolution helps in accurately capturing the details of the characters and improving the performance of OCR systems. Figure 1 represents the handwritten font style of some characters in the CEDAR character set.

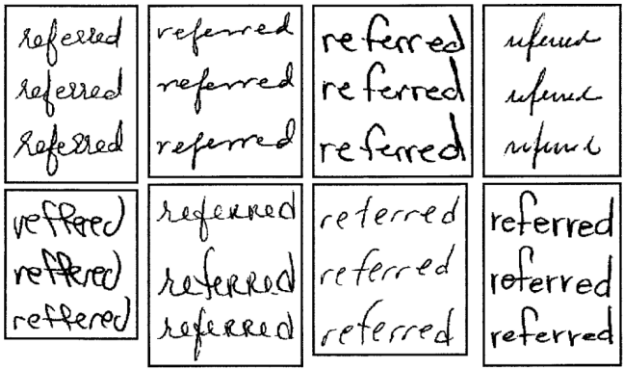

Figure 1 Sample handwritten in CEDAR.

3.2.2 MNIST

This dataset is deemed to be one of the most used/quoted handwritten digital datasets. The data set comprises 60,000 training images and 10,000 test images that have undergone normalization to 20x20

grayscale and resizing to 28x28. This simplifies the pre-processing and formatting process greatly, saving significant time.

3.2.3 HCL2000

This is a handwritten Chinese character dataset. The dataset is comprised of 3,755 frequently used Chinese characters, obtained from 1,000 separate objects. What sets it apart is the presence of two sub-datasets - one for hand-written Chinese characters, and another for the accompanying writer details.

3.3. Model

This chapter provides a classification of the diverse recognition methods used in OCR technology. It provides a comprehensive understanding of the different techniques, their advantages, and limitations. The classification is aimed at giving a clear picture of the various methods used in text recognition.

3.3.1 Artificial Neural Networks(ANN)

Artificial neural network technology is now used in many technologies. It is inspired by biological neurons and stimulates the activity of neurons, which are modeled by inputting the corresponding data and then mapping them to predefined labels. In neural networks, nodes are the basic units or neurons in analog biology. By continuously making inputs, the weights of each node are mediated, thus reducing the error and making the learning result of this neural network closer and closer to the correct result.

Different kinds of Artificial Neural Networks have been developed so far, such as Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN). These techniques have also been applied to OCR technology by different researchers, and thus different features have been developed.

OCR using Convolutional Neural Networks (CNN) has been reported to be a great success, and many commercial OCRs have started to adopt systems using this method. With the powerful learning ability of neural networks, OCR is now widely used in almost all languages around the world.

Recently, researchers have started developing OCR that combines RNN and CNN neural networks. The network transcribes a sequence of convolutional features to convert an input image into a set of target labels. This removes the need to break down the image into individual characters/glyphs, which was a challenging task in the past, especially for scripts like Urdu.

Moreover, it is obvious that increasingly researchers, while adopting the core concept of RNN or CNN, have started to integrate some of their own algorithms to solve some of the problems that have arisen in traditional CNN-based or RNN-based OCR, i.e. how to solve the problem of character recognition accuracy in difficult languages such as Urdu, and the results of these researchers have been able to improve the recognition accuracy of this language up to 99% [15].

3.3.2 Template Matching

The template matching method is arguably one of the longest-standing OCR methods used to date [6]. The principle of this method is easy to understand, as it only requires processing the swept image, extracting a small portion of the image, and then matching it with a predefined template. If the match is successful, the character can be recognized. Usually, template matching uses a sliding window method to match the images in the slider to determine their similarity, which improves efficiency and greatly increases the accuracy of recognition.

Earlier this method was basically applied to recognize standard font characters because of the high accuracy of template matching for standard characters. Nowadays, some researchers have started to apply this method to handwritten characters as well, and the most common method is deformable template matching.

Another method commonly used for template matching method is rigid template matching. this method often extracts the image features first and then performs recognition. And the most widely

used extraction method is the Hough transform, which is often employed for recognizing characters with unique features, such as those in Chinese and Arabic scripts.

3.3.3 Structural Pattern Recognition

It is to extract the primitive information of characters by features such as edges, connections, contours, and combinations of characters, and then to recognize the extracted information for processing.

One of the most commonly used primitives in this approach is the Chain Code Histogram (CCH). It can effectively split characters into different edges to help classify them. CCH is generally not effective for recognizing handwritten characters as the irregular edges of handwritten characters make recognition more challenging, causing significant inaccuracies in recognition outcomes [16]. OCR research categorizes structural models into two groups based on the structure used: graphical models and grammar-based models.

4. Conclusion

OCR technology has been around for nearly eighty years since its inception in the last century. Initially, the recognition and speed of OCR are very low because of the backward algorithm and the low computing speed of the computer. But with the gradual development of the technology, the algorithm has been optimized, the model has been changed, and more advanced machine learning and neural networks have been used in OCR.

Researchers naturally came up with the idea of recognizing standard characters, so most of the OCR technologies produced in the last century could only recognize some pre-defined characters, and these characters in turn had high requirements for their fonts, and the languages recognized were basically single, with English as the main language. With the adoption of different methods, such as template matching, Structural Pattern Recognition, and later neural networks, the accuracy of OCR recognition has been effectively improved and the speed of recognition has leaped exponentially compared to before. The recognition of languages has been extended from individual languages to almost all languages in the world, especially for languages that are very structured, such as Chinese and Urdu.

Since the recognition of standard fonts has reached a high level, the researchers started to focus on the recognition of handwritten characters, and from this point on, they improved the methods previously used in OCR for standard font recognition, discarding some algorithms that are not suitable for OCR, such as Structural Pattern Recognition.

Today, OCR technology is adapted to almost all commonly used languages around the world, especially for the recognition of handwritten characters. It is not difficult to find that OCR has also matured in commercial applications, and modern cell phone cameras are equipped to implement scan recognition. More researchers are beginning to work on future research directions for OCR.

In the future, the current trend is that more researchers will invest in using OCR technology for the recognition of fragmented characters, and this work also shows the maturity of OCR that relies on neural networks. There is already research into the use of OCR for the repair and recognition of fragmented documents, and it is possible that this technology could be used in the future for the restoration of old, damaged books.

References

[1] T. Sakai, M. Nagao, and Y. Shinmi, “A character recognition system,” in Proc. Annual Conj IECE Japan, 1963, p.,450.

[2] M. H. Glauberman, “Character Recognition for business machines,” Electronics, pp. 132-136, Feb. 1956.

[3] W. J. Hannan, “R. C. A. multifont reading machine,” in Optical Character Recognition. G. L. Ficher et al., Eds. McGregor & Wemer, 1962, ip.3-14.

[4] ERA, “An electronic reading automaton,” Electronic Eng., pp. 189-190, Apr. 1957.

[5] R. W. Weeks, “Rotating raster character recognition system,” AIEE Trans., vol. 80, pt. I, Communications and Electronics, pp. 353-359, Sept. 1961.

[6] T. Sakai, M. Nagao, and Y. Shinmi, “A character recognition system,” in Proc. Annual Conj IECE Japan, 1963, p. ,450.

[7] C. Y. Suen and S. Mori, “Standardization and automatic recognition of hand-printed characters,” in Computer Analysis and Perception, vol. 1, Visual Signals, C. Y. Suen and De Mori, Eds. Boca Raton, FL: CRC Press, 1982, pp.41-53.

[8] A. Krzyzak, W. Dai, and C. Y. Suen, “Unconstrained handwritten classification using modified backpropagation model,” in Proc. Frontiers in Handwritting Recognition (CENPARMI Concordia University), 1990, pp. 155-164.

[9] J. J. Hull et al., “A blackboard-based approach to handwritten Zip Code recognition,” in Proc. US. PostalServiceAdv. Techol. C&f, 1988, pp. 1018-1032.

[10] S. Lavirott and L. Pottier, ‘‘Mathematical formula recognition using graph grammar,’’ Proc. SPIE, vol. 3305, pp. 44–52, Apr. 1998.

[11] H. Zhang, J. Guo, G. Chen, and C. Li, ‘‘HCL2000—A large-scale handwritten Chinese character database for handwritten character recognition,’’ in Proc. 10th Int. Conf. Document Anal. Recognit. (ICDAR), 2009, pp. 286–290.

[12] D. Lin, F. Lin, Y. Lv, F. Cai, and D. Cao, ‘‘Chinese character CAPTCHA recognition and performance estimation via deep neural network,’’ Neurocomputing, vol. 288, pp. 11–19, May 2018.

[13] Chaudhury, Ayan, et al. "A Deep OCR for Degraded Bangla Documents." Transactions on Asian and Low-Resource Language Information Processing 21.5 (2022): 1-20.

[14] C. Moseley, Ed., Atlas of the World’s Languages in Danger. Paris, France: UNESCO Publishing, 2010.

[15] Naseer, Asma, and Kashif Zafar. "Meta features-based scale invariant OCR decision making using LSTM-RNN." Computational and Mathematical Organization Theory 25 (2019): 165-183.

[16] Y. Akbari, M. J. Jalili, J. Sadri, K. Nouri, I. Siddiqi, and C. Djeddi, ‘‘A novel database for automatic processing of persian handwritten bank checks,’’ Pattern Recognit., vol. 74, pp. 253–265, Feb. 2018.

[17] H. S. Baird, “Feature extraction for hybrid structural/statistical pattern classification,” CVGIP, vol. 42, pp. 318-333, 1988.

[18] C. Nadal, R. Legault, and C. Y. Sum, “Complementary algorithms for the recognition of totally unconstrained handwritten numerals,” in Proc. IOth IJCPR, June 1990, pp. 443-449.

[19] Bai, Zhen-Long, and Qiang Huo. "A study on the use of 8-directional features for online handwritten Chinese character recognition." Eighth International Conference on Document Analysis and Recognition (ICDAR'05). IEEE, 2005.

[20] Hwang, Wonseok, et al. "post-OCR parsing: building simple and robust parser via BIO tagging." Workshop on Document Intelligence at NeurIPS 2019. 2019.