OPTICAL CHARACTER RECOGNITION TECHNIQUES: A Review

Shubham Srivastava
Electronics & Telecommunication Dept.
SGSITS Indore
shubham7687@gmail.com

Ajay Verma
Electronics & Instrumentation Dept.
IET DAVV Indore
averma@ietdavv.edu.in

Shekhar Sharma
Electronics & Telecommunication
SGSITS Indore
shekhar.sgsits@gmail.com

2022 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS) | 978-1-6654-1418-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/SCEECS54111.2022.9740911

***Abstract*- Optical Character recognition is an eminent topic of research in modern times. An exhaustive research is going on character recognition of different languages. These languages mainly include English and Devanagiri in India. A lot of research has been carried out already in the English language. Devanagiri consists of over 120 regional languages, which is a current topic of research nowadays. Character recognition is done on two types of documents typed and handwritten. In case of Indian and other languages OCR systems are not yet able to recognize the characters successfully with 100 percent reliability because of variation in scripts, quality, size, font, and style. Various algorithms are being developed now days for increasing the reliability of these characters for accurate recognition.**

***Keywords*-Character Recognition, Document Analysis, Optical Character Recognition.**

I. Introduction

Optical Character Recognition finds its applications in the field of Pattern recognition where the main aim is to recognize different characters or numerals in various languages. Character recognition finds many applications in banks, postal services, forensic research, analysis of old documents, etc. In the applications mentioned above, the need is to recognize the characters present in various types of documents such as historical manuscripts and printed scripts. Scripts are mainly classified on the basis of different languages as typed and handwritten and thus the recognition can be classified in two ways as- Printed Character Recognition (PCR) and Handwritten Character Recognition (HCR). Printed Character Recognition is comparatively easier as compared to Handwritten Character Recognition because the available fonts are very limited in number in case of PCR as compared to different diversified handwriting styles in HCR. Handwritten character recognition (HCR) can also be classified into two categories- Offline Recognition and Online Recognition. Recognition in online mode is done in real-time i.e., at the time when the user is writing the document, whereas already written documents are recognized in Offline mode.

II. Approach

Character recognition is done using sophisticated systems and hardware such as Computers. The whole process of character recognition consists of 6 steps. They are mainly classified as (i) Image Acquisition (ii) Pre-processing (iii) Segmentation (iv) Feature Extraction (v) Classification (vi) Recognition. A brief overview of these processes is as follows:-

(i) Image Acquisition

The first and foremost step in the process of character recognition is image acquisition. This task is performed by scanning the images of the documents for which character recognition is to be done [1]. These set of images is also called as Dataset. Generally, there are plenty of sources where these data sets are available in different formats, such as on Internet or in any Central Library of an organization. These formats may also contain very old documents. These documents may contain printed characters or handwritten characters. An Image Acquisition process may also be considered as the digitization of the text available on these documents. Digitization of these documents is useful in many ways as one can easily access them anywhere anytime in today's world.

(ii) Pre-processing

Pre-processing is another important step in the process of character recognition. In pre-processing, different properties of an image is usually altered for various reasons such as dimensionality reduction for fast computation, noise removal for increasing recognition accuracy, image resizing, skew correction, image thinning, image binarization and many more [1]. Generally, median filtering is used for the noise removal in images.

(iii) Segmentation

Segmentation in images is done after pre-processing. In images, segmentation is generally based on the level at which the recognition is to be done. These levels can be categorized into four types (i) paragraphs (ii) lines (iii) words and (iv) characters. Various methods are used for segmentation of documents at various levels such as Vertical Projection Profile, Blob Analysis, etc.

(iv) Feature Extraction

Feature Extraction is one of the key aspects concerned with recognition. Some of the basic feature extraction techniques include Texture features, Structural features and conventional transforms like Fourier Transform and Wavelet Transform. These transforms extract the fine details present in the structure of the characters mainly useful in recognition of Devanagiri characters. Texture features are script independent i.e., these features can be applied to different scripts for their identification especially for characters with smooth contours and randomly oriented edges that are present in the text of some specific languages such as in Arabic and Hindi scripts.

(v) Classification

The problems of classification and their different approaches of solution forms one of the fields of machine learning. There is a wide range of algorithms in the area of classification that shares one single and common objective, but from different views and perspectives. Various classification methods can therefore be categorized on the basis of three different criteria's which depends on the type of learning i.e., unsupervised, supervised and semi-supervised. The goal of learning is to build a model which consists of different class labels in terms of different feature vectors. Some of the basic classifiers used for classification are (i) k-NN classifiers, (ii) Support Vectors Machines (iii) Naive Based classifiers, etc.

(vi) Recognition

Recognition in OCR is a process of identification of characters of different languages. The main aim of recognition is to recognize the character with maximum accuracy. The figure below shows the block diagram which describes the step by step approach discussed above. Although some of the standard datasets which are available online are extensively used for research purpose and extremely high accuracies are obtained on these datasets. But there are many other datasets on which less accuracy are obtained because these datasets may contain different handwriting styles of

978-1-6654-1418-0/22/$31.00 ©2022 IEEE

the same characters. Therefore, one should approach in the direction of developing algorithms which are script independent or structure independent.

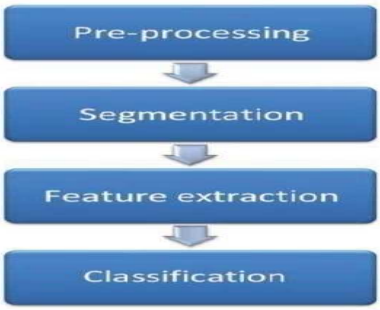

Figure1:- Methodical Approach for OCR

III. Literature Review

In this section we will discuss about some researches done in this field. Some of the notable works are:-

In case of cursive English handwriting each character in a word is connected to each other which create problems during the segmentation of characters, also each person has different handwriting styles. To deal with these issues different types of projection (vertical and horizontal) methods are used. These issues are addressed by Pritam et al.[2], in this paper proposed an orthogonal projection method which is used to normalize skew angle, for feature extraction convex hull algorithm is used and high accuracy is achieved using SVM classifier. An accuracy of 92% is achieved using the above method.

For manually handwritten text's identification, various formats and textual styles can be used by improving the structure of the traditional ANN. "The result of the text is then fed into OCR for conversion of typed, printed or handwritten text into machine-editable text. In this paper offline handwritten character recognition system has been introduced without feature extraction and it was able to achieve 90.19% accuracy."[14] It is used for the recognition of English alphabets.

Shubham et al. [15] proposed a new solution to recognition by using the methods of vision and deep learning for traditional handwriting. It uses EMNIST dataset which is an extension of MNIST digit dataset. Conversion of handwritten text into virtual form is done using CNN for an android application. Preprocessing of EMNIST dataset is proposed by applying different filtering algorithms on it. This model is prepared using the Android studio and links their idea of handwritten text recognition into it. The proposed method also worked on Keras graph which uses tensor flow interfaces and protobuf file to predict the alphanumeric characters generated using a finger. The accuracy achieved using the proposed method is 87.1%.

South Indian scripts include mainly Kannada, Telegu, and Malayalam. These languages are some of the popular languages of India. V. N. Manjunath et.al [3] in this paper uses Fourier Transform and Principal Component Analysis (PCA) for recognition of these languages. This algorithm recognizes all the basic characters of South Indian scripts (lowercase, uppercase) with numerals providing an accuracy of 95.1%.

Gujarati is one of the Indian languages spoken by over 40 million people in India. Gujarati is among some of the popular languages spoken in India. Jayshree et.al in this paper describes the concept of a unique pattern descriptor and Gabor phase XNOR pattern as two newly proposed features which are used for the isolation of handwritten Guajarati characters. Other than the addition to these, two features named as contour direction probability distribution function and autocorrelation features are also proposed. Another useful contribution in this literature is the weighted k-NN classifier and novels mean χ^2 distance measure [4]. In this paper, the classifier uses this new proposed distance measure along with a triangular distance and Euclidean distance to improve the performance of conventional k-NN classifier. The implementation of the above model on a given dataset provides an accuracy of 86.33%.

Obaid et al. [16] proposed a model in which a three layered Artificial Neural Network (ANN) is used for recognition. This paper uses a bitmap representation of an input image which is used as feature vector for the purpose of recognition. An accuracy of more than 95% is achieved by the trained system with test images. The Learning algorithms used in this paper are – 1. ANN Train and Test, 2. Scaled conjugate gradient and 3. Resilient Back Propagation. These Learning algorithms are able to, provide tilt correction, do recognition of text of different sizes, styles and varying background and achieve high accuracy.

Olarik Surnita et.al proposed handwritten character and digit recognition for three scripts of different languages. These languages are Thai, Bangla and Latin. In this paper different local gradient feature descriptors named as histogram of oriented gradients, key point descriptor and scale invariant feature transforms are used. The main aim is to make the system to automatically extract and recognize the characters of the above mentioned languages that are present in the manuscript. The most important point which should be emphasized here is that the quality of the document is one of the factors which can improve the recognition accuracy. It is very essential to take care of the problems which arise during the scanning of these documents. Some of the main problems are distortions present in the image of characters, noise of the background which generally appears at the time of scanning, some amount of skewness present in the documents after scanning. In this paper, MNIST dataset is used and the feature vectors are extracted from the dataset containing handwritten images using local gradient feature descriptors which are then given as an input to a machine learning algorithm to perform the actual classification [5]. For character classification, algorithms such as Support Vector Machine (SVM) and k-nearest neighbor (k-NN) are used. Support Vector Machine gives better results as compared to k-NN. It achieves a highest accuracy of 98.93% on Thai dataset.

In [17], D. H. Kulkarni et al. developed the concept of cloud based intelligent parking services for the Metro cities in India using Internet of Things (IoT) to implement a Number plate recognition system. This method proposed an algorithm that effectively uses the smart parking model and integrates a set of various schemes working on the technology of Internet of Things (IoT). This system makes use of an Ultra-sonic sensor to keep a watch on the car parking. [19]. The different devices could be monitored, tracked or prohibited with the help of computers connected using the Internet.

The Gurumukhi script consists of ten numerals and thirty-five distinct characters. The Gurumukhi script is cursive and there is no format of lower or upper case characters, but many characters in Gurumukhi script are similar in shape which makes recognition of these characters difficult. A combination of Vertical and Horizontal Projection Feature Extraction technique is proposed in which projection histograms are introduced [6]. Thus in this technique the projection histograms gives the count of the number of background and foreground pixels in a particular direction and vertical, horizontal traversing of the normalized binary character image is used for feature extraction.

Classification is done with the help of Support Vector Machine (SVM) and k-Nearest Neighbor (k-NN). By using the combined horizontal and vertical projection technique a maximum accuracy of 97.94% is achieved.

Hasnain Raza et al. [18] discussed the optical character recognition in an application of examination process control. In this paper, various details of a student such as enrollment number, marks, signature, etc is extracted from the examination answer script by segmentation. Enrollment number and mark is used for classification using Convolution Neural Network. Feature extraction from the signature is done using SIFT method. The accuracy of this model is 76%.

Texts present in many scenes when read by automatic machines are mainly restricted to low character recognition accuracy. Shangxuan Tian et al. [7], proposed two new feature descriptors obtained by extending the concept of Histogram of Oriented Gradient Technique (HOG), they are Co-occurrence HOG (Co-HOG) and Conventional Co-HOG (ConvCo-HOG) for accurate recognition of texts of different languages in many scenes. As compared to HOG (which counts the orientation frequency of each single pixel) and ConvCo-HOG (which extracts Co-HOG features from different patches of a character image for more spatial information) [11], the Co-HOG encodes more spatial contextual information by finding the co-occurrence of orientation pairs of neighboring pixels. The scene images consist of texts of different languages of Bengali, English and Chinese. It provides a maximum accuracy of 81.7%.

Raghavendra S.P and Ajit Danti. [20] uses Feed Forward Neural Network and Binary Pattern method for recognition or identification of Bank Cheque Names . In this paper, a Binary Pattern method is used for the segmentation of the bank name and Artificial Feed Forward network is used for the classification of various bank names. [21] Names of different bank cheques are recognized with an accuracy of 92.02%.

In image analysis, it is always very challenging to classify texture images with different orientations and scale changes. Log-polar wavelet signature is an efficient scheme for rotation and scale invariant feature extraction, but by applying Log-polar transform a new row shifted image is produced. In order to eliminate these shift effects, the image is then passed to an adaptive row shift invariant wavelet packet transform. So the output wavelet coefficients are rotation and scale invariant [8]. Chi-Man Pun et al. in this paper describes about the computation of Log-polar wavelet energy signatures which are used for a scale and rotation invariant texture classification and is generally extracted from each sub band of wavelet coefficients. The mentioned approach is tested on different types of data sets and an accuracy of 90.8 percent is achieved for joint rotation and scale invariance.

A novel method for discrimination of languages by analysis of texture properties of the text containing images is proposed by Darko Brodik et al., in this method there is a predefined script type and mapping of each letter is done with this script type, also the position of the character with respect to the baseline area is determined which may be very useful in identification of the characters of different languages. This particular characteristic is used to extract features for characters of different languages. It may be considered as a unique characteristic of different languages. To extract features, the co-occurrence matrix is computed and texture features are calculated. In this paper various statistical parameters are calculated for recognition, after that the texture features are calculated. All these extracted features, shows considerable and meaningful [9] differences because of the different dissimilarity present in script appearances and language characteristics. So one can conclude that, the appearance of any script and its language characteristics plays an important role in the process of its identification. Feature classification for document analysis is done by using one of the soft computing techniques called as Genetic image clustering algorithm. This method is tested on documents containing various scripts such as French, Serbian, English and Solvenian languages. The results of the proposed method are superior to the existing methods.

Extensive research on identification of handwritten numerals and characters of different languages is carried out in recent years. Different types of methods are available for pre-processing, feature extraction, post-processing and classification. Some issues which arise during the development of a handwriting recognition system are feature set selection and design of a classifier. Some of the classifiers [10], include neural classifiers such as radial basis function (RBF), multilayer perceptron (MLP), statistical classifier such as the modified quadratic discriminant function (MQDF), support vector machine (SVM), learning vector quantization (LVQ) classifier, polynomial classifier (PC). Many classifiers such as RBF, MLP, PC, MQDF and LVQ have higher efficiency with respect to both computational and memory cost.

Amit Choksi, Kajal Kumari et al. [22] describes in this paper about the Printed Hindi Character Recognition in documents by analyzing the k-Nearest Neighbor calculation. The authors in this paper develop various preprocessing, division and highlight extraction techniques. A k-nearest neighbor classifier (k-NN) is used to characterize the Hindi character.

Deep Learning based approaches are also being used to recognize characters of different languages. Research is going on to teach the machine to generate these characters automatically [13]. Recently, one of the models known as Recurrent Neural Network (RNN) is used for both recognition and generation of Chinese characters. RNN uses a combination of Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) on the ICDAR-13 model. RNN uses a conditional generative with the character embedding to draw recognizable Chinese characters.

Zhe Li et al. [23] in this paper discuss about the online recognition of Handwritten Chinese alphabets using Two CNN algorithm. Feature extraction is done using the Path Signature method. The accuracy of recognition achieved in this method is about 97.38% on both ±45◦ non-rotated and rotated characters.

Museums, libraries and various other organizations have large collections of handwritten historical documents, for example, the Library of Congress contains paper documents of early presidents like George Washington. The first step for character recognition is to provide recognition of different tools to automatically segment handwritten pages into words. Gap metrics is among one of the most developed and tested algorithms which is used on highly constrained documents like postal addresses and bank cheques. In full handwritten documents, characters are sometimes very difficult to recognize because these documents are needed to clean before recognition. Historical manuscript images, generally contains a large amount of noise and therefore are much more challenging to recognize. Hence, these documents require filtering at the initial stage. In this paper, a novel scale space algorithm is discussed which tries to explain the procedure for an automatic segmentation of handwritten (historical) documents into words. In this process, the first step involves the cleaning of the document to remove all

margins [12] and then a gray-level projection profile algorithm is used which identifies lines in images. All these lines present in the images are then filtered at several scales using an anisotropic Laplacian filtering method. This technique produces a blob which includes words at larger scales and character at small scales. The scale should be selected such that these blobs correspond to words. These blobs are bounded by rectangular boxes to recover the words. This technique outperforms the gap-metrics word segmentation algorithm. The table below briefly compares the recognition accuracy of characters of different languages using various methodologies.

S.No.	Languages	Methodology	Accuracy (%)
1	Cursive English Handwriting	Support Vector Machine (SVM)	92
2	English Alphabets	Artificial Neural Network (ANN)	91.5
3	English Alphabets	Convolutional Neural Network (CNN)	87.1
4	South Indian Scripts	Fourier Transform and Principal Component Analysis (PCA)	95.1
5	Gujrati Scripts	Weighted k-Neural Network (k-NN)	86.33
6	English Letters	Three-Layer ANN	95
7	Thai Dataset	Support Vector Machine (SVM)	98.93
8	Gurmukhi Scripts	SVM and k-NN	97.94
9	English, Chinese, Bengali	Co-HoG and ConvCo-HoG	81.7 71.0 92.2
10	Devanagiri Numerals	Three-Layer Multi-Layer Perceptron	99.04
11	Printed Hindi Characters	Fuzzy k-NN	98
12	Generated Chinese Characters	Recurrent Neural Network (RNN)	93.98
13	Chinese Handwritten Alphabets (+/- 45^0 Rotation)	Two-Stage CNN	97.38

Table1. Recognition Accuracy of Characters of Different Languages Using Various Methodologies

IV. Conclusion

In this paper, we have reviewed some general approaches for character recognition. The different approaches discussed in this paper are applicable to both Printed as well as Handwritten, character recognition. These approaches also work on the documents containing multilingual characters. From the above table, one can conclude that a fairly good percentage of accuracy is obtained for English alphabets (95%) using a three layer ANN, Chinese alphabets (97.38%) using two-stage CNN and Printed Hindi characters (98%) using Recurrent Neural Network (RNN). However, Cursive English letters have a low percentage accuracy of 92% and hence demands for developing or using efficient algorithms for increasing accuracy. Similarly, Guajarati scripts also have low accuracy percentage and thus demands the same treatment as required by cursive English letters. One can also conclude that a very little amount of work is done on handwritten character recognition of Devanagiri languages, especially on Hindi. The recognition rate of Hindi language is generally not very high because of the following reasons:

(i) The complex structural pattern of the characters.
(ii) Presence of Shirorekha.
(iii) Presence of conjuncts in the language.
(iv) Presence of signs (matras) in the language.

Another important issue in the process of recognition is a very close similarity among some of the characters in the Hindi language for example bha and ma, ya and tha because it will reduce the recognition rate. Besides this the character recognition accuracy also depends on the quality of the material which is to be recognized, because an unwanted sign, roughness, oil patches, variation in the ink intensity used in the document, variations in the handwriting of persons also affects it. After the study of an extensive research one will come to an important conclusion that the choice of features to be extracted and the method used for classification plays an important role in performance of character identification rate.

References

[1] B M Vinjit, Mohit Kumar Bhojak, Sujit Kumar, Gitanjali Chalak. A Review on Handwritten Character Recognition Methods and Techniques. *International Conference on Communication and Signal Processing*, pp.1224-1228.

[2] Pritam Dhande, Reena Kharat, "Recognition of Cursive English Handwritten Characters", IEEE, *International Conference on Trends in Electronics and Informatics ICEI 2017*, 11-12 May 2017, Tirunelveli, India, 2017.

[3] V.N. Manjunath Aradhya, G. Hemantha Kumar, S. Noushath Multilingual OCR system for South Indian scripts and English documents: "An approach based on Fourier transform and principal component analysis", *Science Direct Engineering Applications of Artificial Intelligence 21 (2008)*, pp. 658–668.

[4] Jayashree Rajesh Prasad, Uday Kulkarni "Gujrati character recognition using weighted k-NN and Mean χ^2 distance measure". *Springer, Int. J. Mach. Learn. & Cyber*. DOI 10.1007/s13042 013-0187.

[5] Olarik Surinta, MahirF. Karaaba, Lambert R.B. Schomaker, Marco A. Wiering, "Recognition of handwritten characters using local gradient feature descriptors". *Engineering Applications of Artificial Intelligence 45(2015)*, pp. 405–414.

[6] Manoj Kumar Mahto, Karamjit Bhatia, R.K. Sharma, "Combined Horizontal and Vertical Projection Feature Extraction Technique for Gurmukhi Handwritten Character Recognition", IEEE 2015 *International Conference on Advances in Computer Engineering and Applications (ICACEA) IMS Engineering College, Ghaziabad, India*, 19-20 March 2015 , Ghaziabad, India, 2015.

[7] Shangxuan Tian, Ujjwal Bhattacharya, Shijian Lu, Bolan Su, Chew Lim Tan, "Multilingual Scene Character Recognition with Co-occurrence of Histogram of Oriented Gradients", *Pattern Recognition*, http://dx.doi.org/10.1016/ j.patcog.2015.07.009.

[8] Chi-Man Pun, Moon-Chuen Lee, "Log-Polar Wavelet Energy Signatures for Rotation and Scale Invariant Texture Classification", *IEEE Transaction on Pattern Analysis And Machine Intelligence*, Vol. 25, No. 5, May 2003.

[9] Darko Brodic, Alessia Amelio, Zoran N. Milivojevic, "Language discrimination by texture analysis of the image corresponding to the text", *Neural Comput & Applic, Natural Computing Applications Forum 2016*, DOI: 10.1007/s00521-016-2527-x.

[10] Ujjwal Bhattacharya, B.B. Chaudhuri, Handwritten Numeral Databases of Indian Scripts and Multistage Recognition of Mixed Numerals, *IEEE Transactions On Pattern Analysis And Machine Intelligence*, Vol. 31, No. 3, March 2009.

[11] Truong T. Nguyen, Hervé Chauris, "Uniform Discrete Curvelet Transform", *IEEE Transaction On Signal Processing*, Vol. 58, No. 7, July 2010.

[12] R. Manmatha, Member, Jamie L. Rothfeder, "A Scale Space Approach for Automatically Segmenting Words from Historical Handwritten Documents", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, August 2005.

[13] Xu Yao Zhang, Fei Yin, Yan-Ming Zhang, Cheng-Lin Liu, Yoshua Bengio, "Drawing and Recognizing Chinese Characters with Neural Network", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40 , issue 4 , April 1 2018.

[14] Attigeri, S., 2018. Neural network based handwritten character recognition system. International Journal of Engineering and Computer Science, 7(03), pp.23761-23768.

[15] Shubham Sanjay Mor, Shivam Solanki, Saransh Gupta, Sayam Dhingra, Monika Jain and Rahul Saxena, 2019. Handwritten text recognition: with deep learning and android. In 2019 International Journal of Engineering and Advanced Technology (IJEAT).

[16] Obaid, A.M., El Bakry, H.M., Eldosuky, M.A. and Shehab, A.I., 2016. Handwritten text recognition system based on neural network. Int. J. Adv. Res. Comput. Sci. Technol. (IJARCST), 4(1), pp.72-77.

[17] D. H. Kulkarni, G. M. Bhingewad, S. P. Shah, T. G. Mahajan, and S. V. Salke, "Smart Parking System and Number Plate Recognition Using Cloud and IOT," International Journal of Innovative Research in Computer, vol. 5, no. 2, Feb. 2017.

[18] M. Rizvi, H. Raza, S. Tahzeeb and S. Jaffry, "Optical Character Recognition Based Intelligent Database Management System for Examination Process Control." International Conference on Applied Science and Technology, pp. 500-507, 2019.

[19] Gurav, R. Gurav, V. Kamble, N. S. Sakhalkar, and S. Mohite, "A Review Paper on Vehicle Number Plate," International Journal of Engineering Research & Technology (IJERT), vol. 8, no. 4, Apr. 2019.

[20] S.P. Raghavendra and A. Danti, "A novel recognition of Indian bank cheque names using binary pattern and feed forward neural network." IOSR J. Computer. Eng. (IOSR JCE), 20(3), pp.44-59, 2018.

[21] Raj, Jennifer S. "A Comprehensive Survey On The Computational Intelligence Techniques And Its Applications." Journal Of Ismac 1, No. 03 (2019): 147-159.

[22] P. A. Choksi, K. Kumari, S. Kanojiya, P. Sahu, and N. Rindani. "Hindi Optical Character Recognition For Printed Documents Using Fuzzy K-Nearest Neighbor Algorithm A Problem Approach In Character Segmentation." Vol.8, pp. 25-34, 2018.

[23] Z. Li, L. Jin, and S. Lai."Rotation -Free Online Handwritten Chinese Character Recognition using Two-Stage Convolutional Neural Network." In 2018 16th International Conference on Frontiers in Handwriting Recognition, pp. 205-210, IEEE, 2018.