ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)
Vol. 2, Issue 3 (July - Sept. 2014)
International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014)

Candidate Path Search Framework for Chinese Text Recognition

[1]P. Ashok Kumar, [II]B. Tejaswi
[I]Student, Dept. of IT, LBRCE, JNTUK University, Mylavaram, Andhra Pradesh, India
[II]Assistant Professor, Dept. of IT, LBRCE, JNTUK University, Mylavaram, Andhra Pradesh, India

Abstract
Handwriting has continued to persist as a means of communication and recording information in day to-day life even with the introduction of new technologies. The changeless development of computer tools lead to the requirement of easier interface between the man and the computer. Handwritten character recognition may for instance be applied to Zip Code recognition, automatic printed form acquisition, or cheques reading. The importance to these applications has led to intense research for several years in the field of off-line handwritten character recognition .Chinese the national language is world's third most popular language after English. Chinese handwritten character recognition has got plenty of application in different fields like postal address reading, cheques reading electronically. Recognition of handwritten Chinese characters by computer machine is complicated task as compared to typed characters, which can be easily recognized by the computer. This paper presents a scheme to recognize Chinese number numeral with the help of neural network.

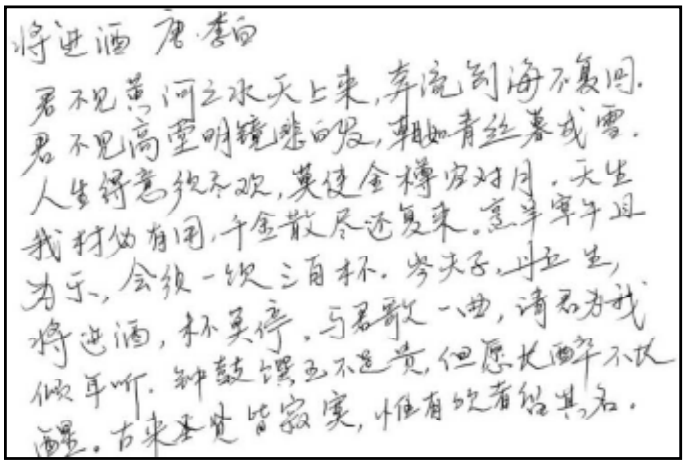
Keywords
Zip code recognition, cheques reading, postal address reading, automatic printed form acquisition, neural network.

I. Introduction

Handwritten Chinese character recognition, it includes the online and offline recognition. Despite of the massive advances and successful applications, there are still remained some of the big challenges, particularly in unconstrained handwriting. Handwritten Chinese character recognition has reported accuracies as high as 98% on sample databases of constrained handwriting but the accuracy on unconstrained handwriting is much lower. Continuous handwritten script recognition is an even more difficult problem. To promote this performance,many efforts are needed to design new methods and databases of unconstrained handwriting are needed for benchmarking. Handwritten Chinese character recognition has been considered as a challenging task. It has seduced much attention since the 1970s and has achieved huge advances [11], [5]. On the other side, many works on online Chinese handwritten text recognition have reported higher accuracies [7], [13], [14], [1]. Online handwriting recognition has the benefit over offline recognition in that the sequences of strokes are available for better segmenting and perceptive characters.

II. Literature Survey

In Chinese character recognition, many methods were assessed on the data sets of the restricted writing styles though very big accuracies have been reported [11]. The accuracy on unconstrained handwritten samples, however, is much lower [2]. In Chinese character string identification, most of the works held at the recognition of text lines or phrases in rather constrained application domains, such as legal amount recognition in bank checks [6] and address phrase Recognition for postal mails [3], [4] where the number of character classes is very small or there are very strong exact constraints. Most of the works on Chinese handwriting recognition of general texts have been reported only in recent years, and the reported accuracies are quite little. For illustration, Su et al. reported character-level correct rate (CR) of 39.37 percent on a Chinese handwriting data set HIT-MW with 853 pages containing 186,444 characters. later works on the same data set, using character classifiers and statistical language models (SLM) rest on over segmentation, reported a character-level correct rate of 78.44 [10] and 73.97 percent [8], respectively.



III. Related Work

There are several issues that need to be solved under the Hand Written Chinese Text Recognition, for example, character over segmentation, character classification, geometric context, linguistic context, path evaluation and search, and parameters estimation. In this section, we briefly outline the importance of linguistic context in HCTR and other NLP tasks as it is the main issue of this paper. A more detailed state of the art on HCTR is available in a comprehensive work recently proposed by Wang et al. After generating candidate character patterns by combining consecutive primitive segments, each candidate pattern is classified using a classifier to assign similarity /dissimilarity scores to some character classes. Character classification needs character normalization, feature extraction, and classifier design. The state-of-the-art methods have been reviewed in. For classification of Chinese characters with large number of classes, the most popularly used classifiers are the modified quadratic discriminant function (MQDF) and the nearest prototype classifier (NPC).

IV. Problem Statement

A. Existing System

In the context of handwritten text recognition, many works have

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

been contributed to the related issues of over segmentation, character classification, confidence transformation, language model, geometric model, path evaluation and search, and parameter estimation about the character recognition. For over segmentation, connected component analysis has been widely adopted, but the splitting of connected characters has been considered as a huge problem. In this the main problem is to recognition of characters which are connected to one another has been a huge problem to recognize the characters. For classification of Chinese characters with large number of classes, the most popularly employed classifiers are the modified quadratic Discriminant function (MQDF) and the nearest prototype classifier (NPC). The MQDF provides higher accuracy than the NPC but suffers from high expenses of storage and computation.

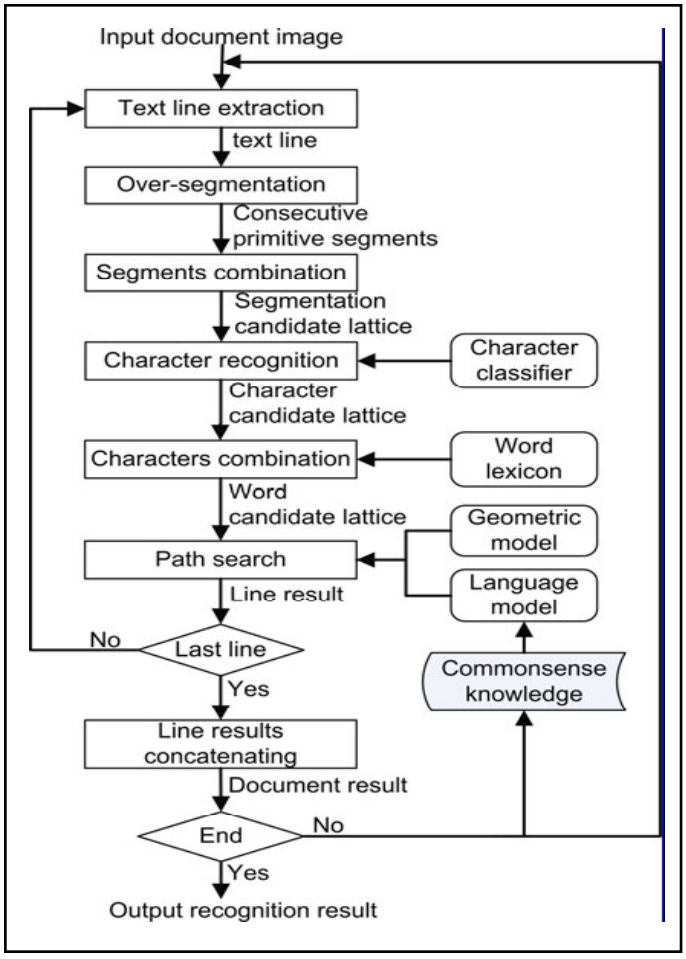


Fig. 1: Architecture of the character recognition

Explanation of the architecture:

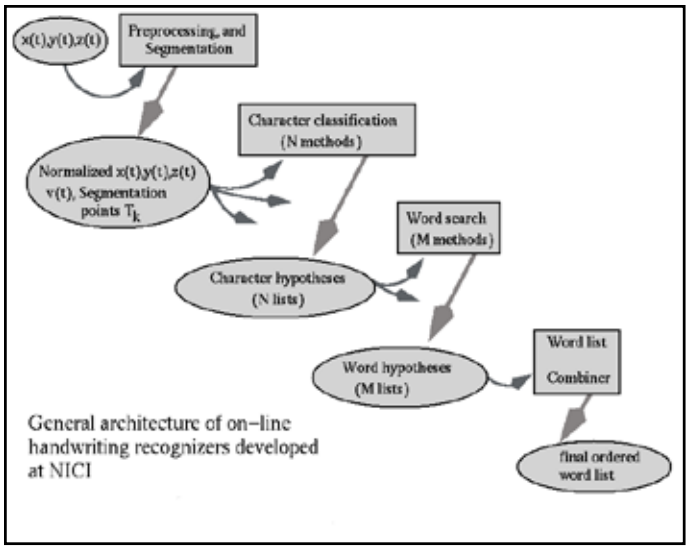
1. Each text line is extracted from the input document image
2. The line image is over-segmented into a sequence of primitive segments and a character may comprise one segment or multiple parts.
3. Several successive segments are combined to generate candidate character patterns where as some are valid character patterns, while some are invalid.
4. Each candidate pattern is classified into several candidate character classes, forming a character candidate lattice
5. Each sequence of candidate characters is matched with a lexicon to segment into candidate words, forming a word candidate lattice
6. Each word sequence C paired with candidate pattern sequence X is evaluated by multiple contexts, and the optimal path is searched to output the segmentation and recognition result
7. All text lines results are concatenated to give the document result, which is used for integrating common sense knowledge in the second pass recognition or output.

Disadvantages

- In this existing system, splitting of connected characters has been a problem for over segmentation.
- In offline handwritten, it is hard to identify text recognition because of different users write their characters in different styles.

B. Proposed System

This system concentrates on the recognition of text lines, which are assumed to have been segmented externally. For the convenience of academic research and benchmarking, the text lines in our database have been segmented and annotated at character level. Now a day's Neural network technique has play a vital role with key of artificial intelligence. It is a strong data modeling tool that is able to capture and represent many variant and connected input/output relationships. The motivation for the evolution of neural network approach originates from the desire to develop an artificial system that could perform "intelligent" tasks similar to those performed by the human brain. The most common neural network model is the multilayer perceptron (MLP). This type of neural network is known as a supervised network because it requires a desired output in order to learn. The goal of this type of network is to create a model that correctly maps the input to the output using historical data so that the model can then be used to produce the output when the desired output is unknown.



The MLP and many other neural networks learn using an algorithm called back propagation. With back propagation, the input data is frequently presented to the neural network. With each presentation the output of the neural network is compared to the desired output and an error is calculated. This error is then feed back to the neural network and used to adjust the weights such that the error decreases with each iteration and the neural model gets closer and closer to producing the desired output. This process is known as "training".

V. System Architecture

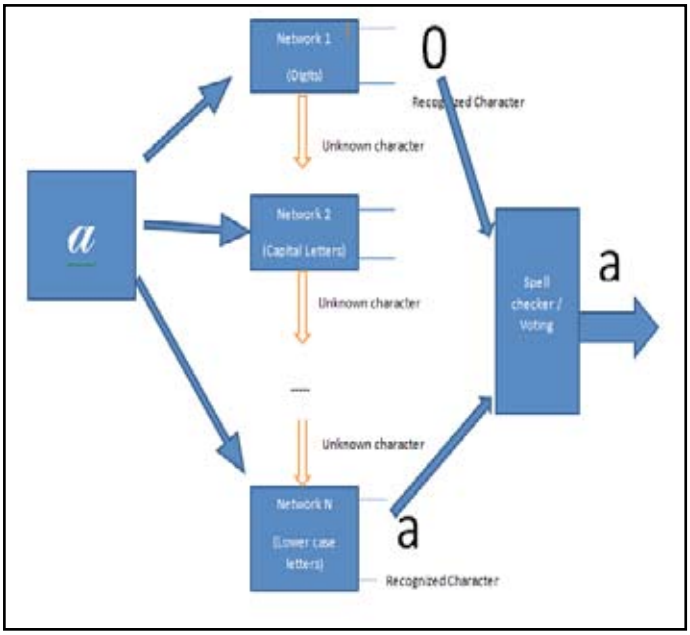


Fig. 2: Network based character recognition

VI. Modules

Modules Description

1. Creation of Draw Panel

This is the first module, where we can create the interface for the users to provide the input texts. The users draw the text in the provided panel area. In case if the text is given wrongly then there is an option available in the panel which is the clear button. It will clears the panel and so that the user can be able to input a new text to recognition.

2. Over-Segmentation

In this module, first the input text line image is over segmented into a sequence of primitive segments using the connected component-based method. Then later the Consecutive primitive segments are combined to generate candidate character patterns, then forming a new segmentation candidate lattice, after that each candidate pattern are classified to assign a number of candidate character classes, and all the candidate patterns in a candidate segmentation path generate a character candidate lattice. If a word level language model is applied, each sequence of candidate characters is matched with a word lexicon to segment into candidate words, forming a word is candidate lattice. All these candidate character lattices are combined to construct the recognition of the segmentation lattice of text line image. Each path in this lattice is developed by a character sequence paired with the sequence of candidate character patterns and this path is called a candidate segmentation recognition path. Finally, the task is to find the optimal path in this segmentation-recognition lattice of the string recognition. Taking that the text lines are segmented from text pages, we utilize the linguistic dependency between sequence of lines is to improve the recognition accuracy by merging the multiples; we get the top-rank recognition results of the previous line to the current line for recognition.

3. Character recognition

Generally the Chinese texts are mix with alphanumeric characters and punctuation marks. And different characters show different outline features, later we design two class dependent geometric models namely, the single-character geometry and the other one is the between-character geometry, which is called as the binary geometric model. In addition, another two class-independent geometric models are designed to indicate whether a candidate pattern is a valid character or not, and checks whether a gap is a between-character gap or not. Then these four geometric models (unary and binary class-dependent, unary and binary class-independent) are denoted as, the unary class-dependent as the "ucg", binary class-dependent as the "bcg" and the unary class-independent as the "uig" and the binary class-independent as the "big" respectively, and have been used successfully in transcript mapping of handwritten Chinese documents.

4. Ranked List Evaluation

We evaluated the effects of different techniques. First, we compared the effects of different path evaluation functions. Second, the sequel of different confidence transformation methods, combinations of geometric models and language models were evaluated. Last, we show the results of different numbers of candidate character classes, beam widths, and candidate character augmentation methods in path search.

5. Result String related definition Module

This module presented an approach for handwritten Chinese text recognition under the character over segmentation and candidate path search framework. The merging weights of path evaluation function are optimized by a string. The experimental results justify that the advantages of the confidence transformation of classifier outputs, geometric context models, and language models. However, the effect of candidate character augmentation is limited. We also evaluated the performance. The aim of the over segmentation is to improve the tradeoff between the number of splitting points and the accuracy of separating characters at their boundaries. The main aim of this character classification is to upgrade the classification accuracy and the tradeoff between the number of candidate classes and the possibility of including the true class.

VII. Experimental Results

Fig. 3: Drawing the character on the canvas

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

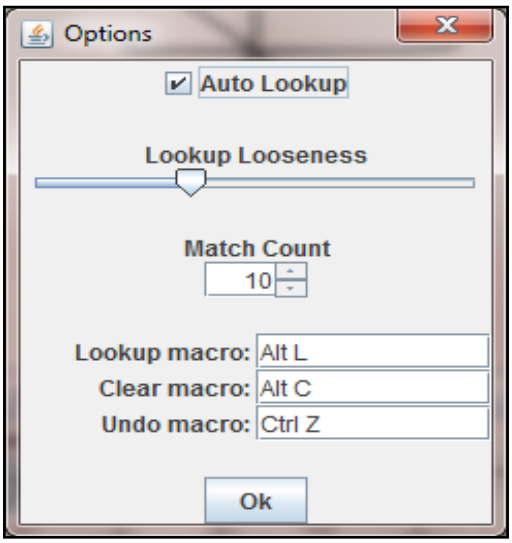


Fig. 4: Lookup options

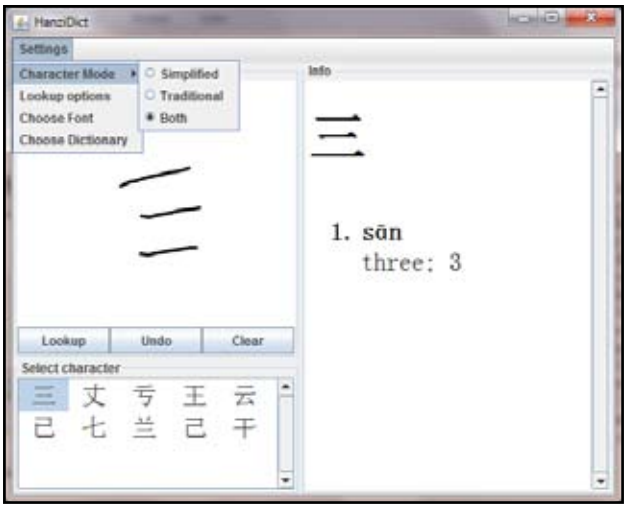


Fig. 5: Character recognition

VIII. Conclusion

This paper presented an approach for handwritten Chinese text recognition under the character over segmentation and candidate path search framework using neural network. We evaluate the paths from the Bayesian decision view by combining multiple contexts, including the character classification scores, geometric and linguistic contexts. In path search, we employ a refined beam search algorithm to refine the accuracy and efficiency. Utilization of neural network approach connected characters splitting problem can overcome on over-segmentation and identification of text recognition as easy task. The desire of over segmentation is to improve the tradeoff between the number of splitting points and the accuracy of separating characters at their boundaries. Our experimental results reviewed that mismatch of language model and text domain induce inferior recognition performance. In addition, the real semantic context and long-distance context will also be considered in this work.

References

[1]. *B. Zhu, X.-D .Zhou, C.-L. Liu, and M. Nakagawa, "A Robust Model for On- Line Handwritten Japanese Text Recognition," Int'l J. Docu ment Analysis and Recognition,vol.13,no.2,pp.121-1312010.*

[2]. *C.-L. Liu, F. Yin, D.-H. Wang, and Q.-F. Wang, "Online and Offline Handwritten Chinese Character Recognition: Benchmarking on New Databases," Proc. Second CJK Joint Workshop Pattern Recognition, Oct. 2010.*

[3]. *C.-L. Liu, M.KogaandH.Fujisawa, "Lexicon - Driven Segmentation and Recognition of Handwritten haracter Strings for Japanese Address Reading," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 11, pp. 1425-1437, Nov. 2002.*

[4]. *C.-H. Wang, Y. Hotta, M. Suwa, and S. Naoi,"Handwritten Chinese Address Recognition," Proc. Ninth Int'l Workshop Frontiers in Handwriting Recognition,pp. 539 544,Oct2004.*

[5]. *H. Fujisawa, "Forty Years of Research in Character and Document RecognitionAn Industrial Perspective," Pattern Recognition, vol. 41,no.8,pp. 2435-2446, Aug. 2008.*

[6]. *H.-S. Tang, E. Augustin, C.Y. Suen, O. Baret, and M. Cheriet,"Spiral recognition Methodology and Its Application for Recognition of Chinese Bank Checks," Proc. Ninth Int'l Workshop Frontiersin HandwrittingRecognition,pp.263-268,Oct.2004.*

[7]. *M. Nakagawa, B. Zhu, and M. Onuma, "A Model of On-Line Handwritten Japanese Text Recognition Free from Line Direction and Writing Format Constraints,"IEICE Trans. Information and Systems, vol. 88no.8,pp.1815- 1822,Aug .2005*

[8]. *N.-X. Li and L.-W. Jin, "A Bayesian- Based Probabilistic Model for Unconstrained Handwritten Offline Chinese Text Line Recognition," Proc. IEEE Int'lConf. Systems, Man, and Cybernetics, pp.3664- 3668, 2010.*

[9]. *Q. Fu, X.-Q. Ding, T. Liu, Y. Jiang, and Z. Ren, "A Novel Segmentation and Recognition Algorithm for Chinese Handwritten Address Character Strings," Proc. pp.974-977, Aug.2006*

[10]. *Q.- F.Wang,F.Yin ,and C.-L. Liu, "Integrating Language Model in Handwritten Chinese Text Recognition," Proc.10thInt'lConf.Document Analysis and Recognition, pp.1036-1040, July 2010*

Authors Profile

P. ASHOK KUMAR received his B.Tech degree in Computer Science and Engineering from Priyadarshini College of Engineering & Technology, Tenali, A.P, India, in 2012. Currently pursuing M.Tech in Lakireddy Bali Reddy College of Engineering, Mylavaram, India. His research interest includes: Image processing and Cloud Computing.

B.TEJASWI received her B.Tech degree in CSE from Narayana Engineering College, Gudur, in 2009 and completed her M.Tech degree in Computer Networks and Information Security from Sree Vidyanikethan Engineering College, Tirupati, in 2013. Currently working as an Assistant Professor in Lakireddy Bali Reddy College of Engineering, mylavaram. Her research interest includes: Cloud computing and Image Processing