Advances in Intelligent Systems Research (AISR), volume 151
2018 International Conference on Computer Modeling, Simulation and Algorithm (CMSA 2018)

Tibetan Character Recognition Based on Machine Learning of K-means Algorithm

Huiwen Gong and Wei Xiang*
Southwest MinZu University, Chengdu, Sichuan, 610225, P.R. China
*Corresponding author

***Abstract*—In this paper, we analyze and extract the Tibetan text features structure based on k-means image character recognition algorithm. Through character library file generated from Tessract-ocr training, we improve the accuracy and recognition of image text recognition and extraction and realize the identification of Tibetan.**

Keywords—artificial intelligence; machine learning; Tibetan character recognition; Tesseract -OCR; K-means algorithm

I. INTRODUCTION

Tibetan language is an ancient phonetic alphabet used by the Tibetan people in China. It has a long history and is recognized as a mature text in the world. Artificial intelligence is a science whose inclusion is very extensive. It is composed of different fields. The field of research including robots, speech recognition, image recognition, natural language processing and expert system, etc. It is of great practical significance to study it. Character recognition [1] is a kind of new technology with the combination of pattern recognition, image processing and word processing technology. It is a concrete research direction in the field of pattern recognition and artificial intelligence [2].It automatically enters text or other information into the computer through intelligent recognition, replacing the manual input process. Under certain conditions, text recognition input is more convenient and faster than keyboard input. After years of exploration and practice, English and Chinese text recognition has been improved. But the technique of recognizing Tibetan printed text is not mature. To solve these problems, according to the latest development of character recognition and extraction, we use an algorithm which is based on k-means image text recognition and extraction to preprocess Tibetan image, process pixel clustering[2],select and optimize layer and cut the text. After testing, the algorithm can be used to identify the Tibetan text effectively.

II. CHARACTERISTICS OF TIBETAN WRITING

Tibetan [3] is a phonetic alphabet which the main component is consonant. Its spelling and structure are different from English and Chinese characters. Tibetan alphabet has thirty consonants and four vowels. The basic unit of a complete Tibetan morpheme is determined by the "syllable delimiter" in Tibetan. A Tibetan word consists of one or more syllables. Each of the horizontal basic units of a syllable is called a character. Each syllable contains the word "base word" and the characters that may be followed, as the former plus word, plus words, vowels, followed by words and so on. Syllables are usually divided by syllables or other punctuation marks. Each syllable contains the "base" and may follow the characters, such as the former plus words, plus words, vowels, followed by words and so on. This is a four-character syllable (FIG. I).

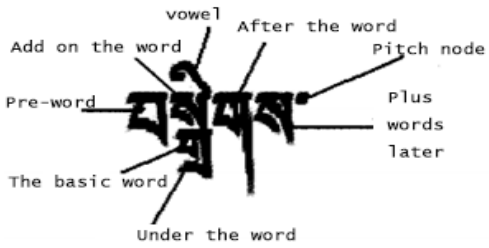

FIGURE I. DECOMPOSITION OF TIBETAN SYLLABLES

III. THE TEXT EXTRACTION METHOD BASED ON K-MEANS ALGORITHM MACHINE LEARNING

A. K-means Algorithm

Literature [5] is based on canny operator detection, using edges and regions to extract text. However, with the advent of the era of artificial intelligence, various algorithms and techniques in the field of big data have been rapidly developed and applied. As a technical field of big data technology, machine learning [4] will play an increasingly important role in big data processing and analysis in the future. Through machine learning algorithm, a meaningful data model is extracted from the mass data in the application development. It can achieve the purpose of discovering data value and realize the realization of data. It predicts future data from the models it obtains. It can also fine-tune the model with future data to better suit the application scenarios.

The basic idea of K-means clustering algorithm is to use the distance between the elements of the set as the dividing standard. Within the collection, the elements are divided into different subsets according to the distribution density of the elements .In the process of division, by defining the distance of elements, we aggregate the elements based on the minimum distance principle between the element and the cluster center to get the final partition result.

An image text recognition and extraction algorithms based on K-means include digital image preprocessing, pixel clustering, layer selection and optimization, and text

segmentation four core steps. It eventually gets a digital image that can be identified by tesseract-ocr.

B. Feature Extraction

1) *Peripheral outline stroke feature extraction:* The commonly used methods are difficult to meet the need for the identification of a variety of printed font Tibetan characters. It is found that the method of extracting features of peripheral contour strokes is better. The method is carried out in four steps.

a) *Peripheral contour extraction:* The purpose of contour extraction is to obtain the external features of the image. The algorithm for contour extraction of binary images is very simple. It is to hollow out internal points and search along the outer outline, right, bottom, and left edges of the contour. Then following the direction of the search, scanning in the opposite direction of the edge point, saving the coordinates of the black pixels until the first black pixels are encountered. Finally, a contour point set T is formed, which is the outline of the contour point. Determine the outline point stroke: Select the Columns icon from the MS Word Standard toolbar and then select1"Column"from the selection palette (Fig.2).

b) *Determine the outline point stroke:* According to the writing sequence of the printed Tibetan characters, the outer stroke of the Tibetan characters is coded in 1-8 directions. We trace the points in the set T of the contour point to obtain the length L of the black pixels connected in the actual pictures in each direction (FIG.III)

c) *To determine the direction of the code string:* The direction code string is to assemble the contours of the same side to form a string.

d) *Amendment coding string:* The entire sample code string will produce a lot of data redundancy after the formation of the code string. Therefore, the code string must be amended and streamlined. The method is: The elimination of sporadic data and the same number of characters, the number of characters is greater than the threshold of the character preservation, on the contrary character removal. Merge the same and connected characters, connected to each other with the same characters, merged into one character.

FIGURE II. STROKE ORDER AND DIRECTION OF STROKE FEATURES

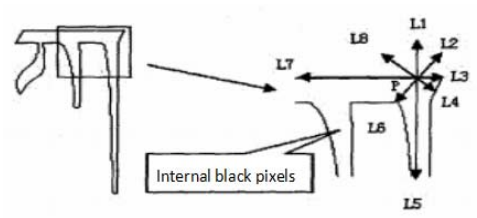

FIGURE III. CONTOUR POINT STROKE

IV. TESSERACT-OCR RECOGNITION ENGINE

Python-tesseract is a Python tool for optical character recognition. OCR is a process of scanning text data, analyzing and processing image files, and obtaining text and layout information. The python-tesseract is an encapsulation of the Google tesseract-ocr. It can also be used as a separate invocation script for the tesseract engine. It supports using the PIL library to read various image file types, including jpg, PNG, GIF, BMP, tiff, and other formats. As a script, it prints out the identified text instead of writing to the file.

Through an analysis method which is based on the mixed page layout based on tabbed bit detection, Tesseract is used to analyze the image layout and distinguish the image form, text, image and other contents. We use k-means image text recognition and extraction algorithm to preprocess digital image, cluster processing pixel, select and optimize layer and cut text. Make image recognition and extract Tibetan text.

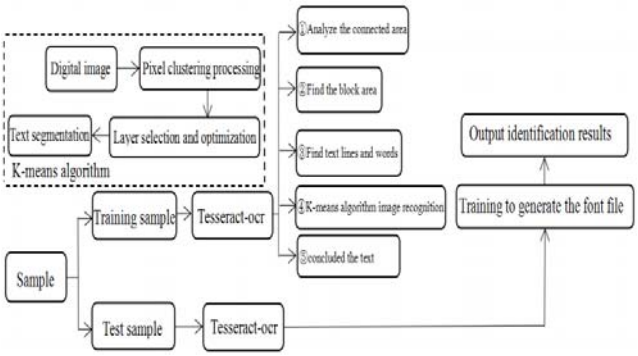

FIGURE IV. OVERALL STRUCTURE OF TEXT RECOGNITION

V. TIBETAN CHARACTER RECOGNITION TRAINING

A. Identification of Tibetan Language Based on Tesseract-ocr Untrained Feature Extraction

Tesseract-ocr is not able to identify the Tibetan language without training. So it is required to train Tesseract.

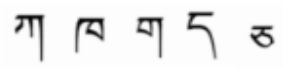

FIGURE V. TRAINING TEXT

1) Training process(FIG.V)

a) Put some pictures in a directory and create a training image tiff.

b) Use the training images generated from the first step to generate the corresponding Box file.

c) Extract character features and generate character feature files. Use the TessBoxEditor to manually correct and save the characters identified in each image.

d) Generate the font file and create a batch file in the directory where the sample image is located.

e) *Merge multiple training files after the font file.*

B. Training Principle Diagram (FIG.VI)

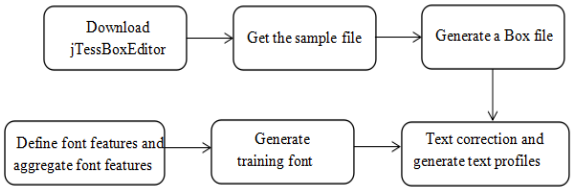

FIGURE VI. TRAINING PROCESS

C. Different Methods Identify Error Rate Comparison (FIG.VII)

TABLE I. K-MEANS ALGORITHM IMAGE TEXT RECOGNITION TEST RESULTS

Text image	Text number	Missing inspection	False inspection	Recall Radio /%	Precision Radio/%	Recognition rate/%
Image 1	15	1	1	94.11	87.11	87.11
Image 2	27	5	2	81.48	82.59	80
Image 3	38	2	2	84.37	84.59	82.73

TABLE II. LITERATURE [5] METHOD TEST RESULTS

Text image	Text number	Missing inspection	False inspection	Recall Ratio /%	Precision Radio/ %	Recognition rate/%
Image 1	15	0	0	100	100	100
Image 2	27	2	2	92.59	92.59	92
Image 3	38	1	2	97.37	94.59	94.73

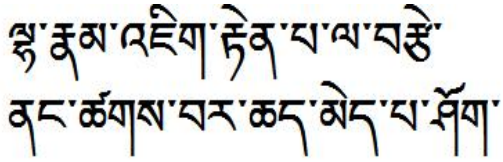

FIGURE VII. TIBETAN EXAMPLE

By investigating and analyzing k-means clustering algorithm and combining with digital image gray scale processing, binarization, edge detection and other technologies, this algorithm can extract and identify the text area in the background of digital image well. Compared with the literature [5], it has higher execution efficiency and accuracy.

VI. SUMMARY

We use an algorithm which is based on k-means image text recognition and extraction to preprocess Tibetan image, cluster process pixel, select and optimize layer and cut text.Through character library file generated from Tesseract-ocr training,we realize the identification of Tibetan. With the continuous development of artificial intelligence technology, people have also become more in-depth research on it. Machine learning is widely used in web search, language recognition and machine vision. In terms of word recognition, Chinese and English technology are mature, but the Tibetan and other ethnic minority language text extraction and recognition are not yet perfect. Through the study of this article, we improve the accuracy of Tibetan recognition. It can be better applied to design Tibetan word recognition software.

ACKNOWLEDGEMENT

This work was partially supported by Sichuan Youth Science and Technology Innovation Research Team (2017TD0028).Also was supported by the Fundamental Research Funds for Central University, Southwest Minzu University (2017NZYQN45).

REFERENCES

[1] Feng Yongkang, Wang Yanhong, Zhang Li. An automatic fuzzy recognition method of text images [J]. Computer Programming Skills and Maintenance, 2015 (15): 82, 89.

[2] Hu Wen, Ma Lingyu, OpenCV mobile phone photo courier single text recognition [J]. Journal of Harbin University of Commerce: Natural Science Edition, 2015, 31 (5): 564-568.

[3] Murray, the analysis of the status quo of minority language translation studies in China [J]. Foreign Language Teaching and Research 2015 Issue 1 P130-140 1000-0429.

[4] Huang Xin, Text Recognition Based on Machine Learning of Artificial Intelligence [J] .Journal World, 2016, (13): 234-234.

[5] Hallidan.A. Yilhamm. Al, kuban. Buy timusha. Segmentation algorithm of uyghur characters in complex background [D]. Computer engineering and application, 2007, 43 (20):163-165.