International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8, Issue-1, May 2019

Handwritten Tibetan Character Recognition using Hidden Markov Model

Nyichak Dhondup, Deepa V Jose

Abstract: ***The Tibetan language which is one of the four oldest and most original languages of Asia is elemental to Tibetan identity, culture and religion and it convey very specific social and cultural behaviors, and ways of thinking. The annihilation of the Tibetan language will have tremendous consequences for the Tibetan culture and hence it is important to preserve it. Tibetan language is mainly used in Tibet, Bhutan, and also in parts of Nepal and India. Tibetan script is devised based on the Devanagari model and Sanskrit based grammars. In this paper, a method for Tibetan handwritten character recognition based on density and distance feature detection is presents. To get a better classification result, images are converted into binary and noise removal is done by using Otzso's method. Features are extracted by normalizing the image based on distance and density of the pixel in the image. Finally, Hidden Markov Model is used for character classification.***

Keywords: ***Tibetan Character, Distance, Density, Otzso, Hidden Markov Model.***

I. INTRODUCTION

Tibetan is a language with history of over 1,300 years; it was created during the time of Tibet king Songtsen Genpo in the 7th century. Currently, more than six million participants use Tibetan language, especially in Tibet. Research on Tibetan language will facilitate the digitization of Tibetan documents, which is very important both from theoretical and practical perspectives which will aid to preserve the rich Tibetan heritage. Handwritten character recognition is gaining increasing interest now a days. However, compared to the existing research work on other languages like Hindi, English, Chinese and Arabic, little work is done on Tibetan handwritten recognition. It is a relatively unexplored field but of great significance and hence the need of this research work. Very few researches have been done related to Tibetan character identification and the efficiency of these algorithms is not satisfactory. Many cases were experimented and found Hidden Markov model (HMM) gives a much better result by solving many segmentation problems and makes classification more accurate.

Revised Manuscript Received on 30 May 2019.
***** Correspondence Author
Nyichak Dhondup*, Department of Computer Science, CHRIST (Deemed to be University), Bangalore, Karnataka, India.
Deepa V Jose, Department of Computer Science, CHRIST (Deemed to be University), Bangalore, Karnataka, India.

So HMM was selected to use with Tibetan characters. For this approach, Otzso's thinning/noise removing and pixel density based segmentation is used. By using the sum of the pixel in a row, getting a hundred percent accuracy of segmentation is very hard and because of that segmentation is still a challenge in this approach as the final accuracy of the system is highly dependent on the segmentation.

Handwritten character recognition can be differentiated into two different categories, Online character recognition and offline character recognition, online handwritten character recognition deals with automatic conversion of character which are written on a system like PC or tablet where sensor picks up the movement of the touch point on the screen, while Offline character recognition deals with a data set, which is obtained from a scanned character or handwritten document. Many researches have been done to recognize handwritten script or character recognition for common language like English, Chinese, Hindi. But Tibetan character recognition using image processing is a completely new field where only a little research had done, and a brief review of methods which were used in those works is mentioned below.

Text line segmentation is a very important and one of the most challenging steps in character recognition, inaccurate segmented text lines will directly affect your final result and Small gaps between nearby text lines cause touching and overlapping of words or letters, as geometrical properties of every words and characters in every text line is different, such as high of the character and the distance between lines. In case of text line segmentation procedure, major difficulties include the difference in the skew angle between lines on the page or even along the same text line[1], to identify the boundary of text lines, they proposed a methods which consist of two steps (i) Generating partial boundary line and (ii) Generating complete boundary line. It is very difficult to differentiate between the two lines by only using these partial lines, thus they generate complete boundary line which acts as differentiator for identifying the text lines and the accuracy of the proposed methods is 98% for handwritten documents of language like Hindi, Kannada, English and Arabic. In the field of recognition of handwritten characters, image zoning is an extensive technique for the extraction of features as it is rightly thought out to be able to manage up with the variability and differences in handwritten patterns[2]. Image zoning is a widely used feature extraction methods which used to obtain information about characteristics of the image, concepts of euler number is used classify the characters. Euler number is obtained by subtracting the number of holes in the image from the number of objects in the image. Long-Long and Jian [3] had proposed a three stage classification strategy to recognize the online unconstrained handwritten Tibetan character.

They used three statistical recognition methods for the feature extraction. They evaluated the recognition performance on two databases of online handwritten Tibetan characters: MRG-OHTC and IIP-OHTC. As a result, the recognition accuracy improved 3% then the existing one. For recognition of Tibetan woodblock print, Fares et al. proposed a two stage framework involving image processing[4], which consisted of noise removal and baseline detection, and character segmentation/ recognition with generalized HMM(gHMM). They constructed a database of 6555 characters from 7 pages of Tibetan wood-block print for experiments. Generalized hidden Markov model resulted in 89.49% accuracy. Later they segmented 104 pages manually and retrained Optical Character Recognition (OCR) on them to see how accurate the performance is. segmentation accuracy increased to 91.29% after use of manual segmentation method. Ahmed and Mohamed [4] proposed three efficient techniques that can be used to discriminate between Arabic and English languages: Peak detection, Moments and Run Length histogram. As a result, they got accuracy performance of 97.8% for Arabic and 98.5% for English based on text line level. These methods can be used for word-level language identification and they got result of 96.1% for Arabic and 98.5% for English. Namsel OCR is a Tibetan OCR which is introduced by Namsel which is an integrated platform generating and disseminating Tibetan electronic text, paper [5] discussed about the basics of how it works, the problem which it can be solved and how the problem solved, and also points a number of areas where it can be improved.

Mallikarjun [6] presented a Gaussian Mixture Model (GMM) to identify the script of handwritten words of Roman, Devanagari, Kannada and Telugu scripts. The authors used a combined approach of Discrete Cosine Transform (DCT) and discrete wavelets Transform (DWT) for feature extraction and neural network (feed forward back propagation) classifier for classification and recognition purpose [7]. Chowdhury et al.[8], proposed an algorithm which can solve the problem of offline character recognition. They had given the input in the form of images. The algorithm was trained on the training data that was initially present in the database. They have done preprocessing and segmentation and detected the line. In[9][10], different kind of methodologies which are used to segment a text based image at different levels of segmentation is discussed and serves as a guide for participants working on the text based image segmentation area of Computer Vision. First, the need for segmentation is justified in the context of text-based information retrieval. Then, the various factors affecting the segmentation process are discussed followed by the levels of text segmentation. Youssef et al.[11] presented a handwritten recognition method based on horizontal and vertical baselines detection features. These features are related to the pixel density and extracted on binary images of characters using the sliding window technique. For the classification, they used multilayer perceptron and experimental results using AMHCD database and demonstrated the efficiency of the proposed system. Atmaprakash et al.[12]presented a Handwritten English Character recognition using Hidden Markov Model (HMM) and Genetic Algorithm to identify features of every characters and compare with characters which are in the testing set, they used various stages of handwritten character recognition system which read a scanned image of handwritten character and converted that into binary form, resizing each matrix into size of n x m where n and m can be any number, and thinning of an image to obtain a clear skeleton of each character. Then they identified each character, using Forward Algorithm. Baum et al.[13] devised a method for the off-line handwritten character recognition using HMM, they compared the performance of the HMM with a baseline of Naïve Bayes classifier.

Volker et al.[14] presented a system using HMM for word recognition based on character recognition without segmentation. Anastasia et al.[15]proposed a HMM system to recognize Japanese characters from a number of different handwritings. They used HMM to optimize the number of state and feature extraction and obtained 95.7% accuracy. Amit et al.[16] described that main recognition accuracy of an OCR system is depended on extracted features obtained by binarization technique for recognition of handwritten characters in English language. The recognition of the handwritten character is done by using multi-layered feed forward artificial neural network as a classifier.

II. PROPERTIES OF TIBETAN CHARACTERS

The Tibetan character set can be differentiate into four vowels and thirty consonants.

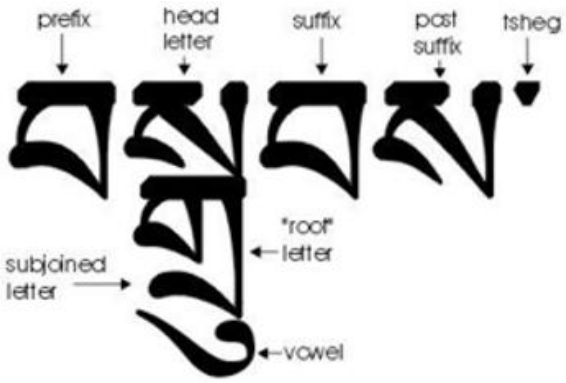

Figure 1. The Syllable Structure

A character can be defined as a arange of consonants with optional vowels. A syllable is a group of characters with one essential consonants and other 7 optional parts, as shown in Figure 1, combine characters consist of at least one essential Consonants, and may include top vowel(TV) or may be

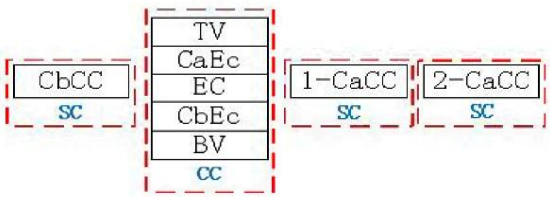

Figure 2: Four-character syllable

bottom vowel(BV), a consonants above the essential consonants (CaEC), a consonants below the essential consonants(CbEC), some consonants can be serve as left side of the essential consonants (1-CaCC), come consonants can be serve as right side of the essential consonants(2-CaCC),

there are few consonants which can be serve at two position right of the essential consonants.
Figure 2 is an example of a four-character syllable. In this research work only single character (SC) recognition is considered.

III. PROPOSED SYSTEM AND METHODOLOGY

The essential goal is to accelerate the process of character recognition in archiving process of Tibetan manuscripts so that current frameworks can handle enormous archives within a short time and can convert them into editable text. Different person have their own different ways of handwriting styles. OCR can recognize the character using optics. Sometimes OCR system cannot understand the same character written by various participants. In order to outdo this, different handwritten symbols written by various participants is collected as shown in Figure 3. The handwritten documents are scanned and treated as images and then trained. Objects which are not matching are excluded and only those images which are matching and having highest result of similarity are taken into consideration, and those images are used for training so that it can understand the same character written in various styles.

A. Data collection and Preprocessing

Handwritten characters appear in various size and shape due to variation in written styles. The data utilized for the experiment is gathered from different participants. There were no restrictions or guidelines in the written style. Same

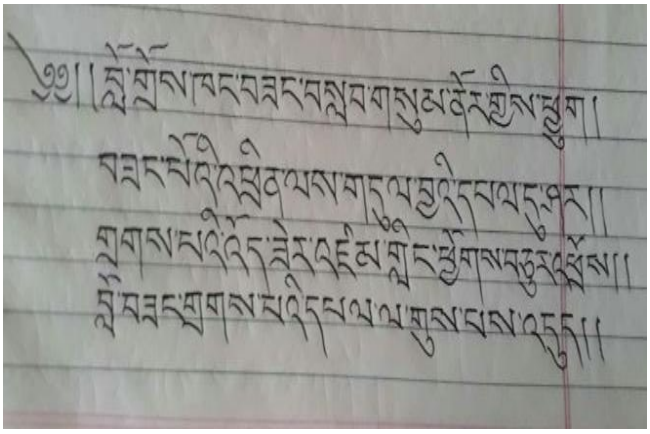

Figure 3: Sample data

content is given to the participants and were asked to write the same in their own style. The writings from 300 participants were collected from monastery, college students and school students. Data which are collected were in the different sizes. A total of 300 handwritten characters are collected and stored as data set. Out of those samples, we used 70% as training data and the remaining 30% to test the recognition classifier. The collected data underwent a certain preprocessing step. First for the segmentation step, binary form of the image is made so that numeral images have pixel values 0 and 1. Then segment the script to the line level and then line level to the character level. For this pixel density in a row to get line level and column to get character is used. After that, Otzso's methods for removing the noise and thinning methods to find the skeleton of the character is performed. Figure 4 shows the character after the preprocessing.

B. Feature Extraction

Features are the representative measures of a signal, which distinguish it from another signals. The selected features should be maximizing the distinguishable features between different Tibetan characters. Features are extracted by transforming the image in time domain to the image in frequency domain, and the find the variation in brightness or color across the image.
To extract features of the character image, two methods are used which is based on pixel density and the distance. As image is in the form of Binary image, adding the sum of pixels in the image, will add only the pixel values of the character part, because the background value of the character is always zero. Then the pixel distance as row and column wise was calculated. Finally, normalize the features which were obtained from extracted image based on density and the distance which lies between 0 and 1 and combined them. The feature extraction algorithms is described in following steps.

Algorithm for feature extraction:
Input: Gray Scale Image.
Output: feature vector computed by performing density and distance-based feature extraction and normalize them.
Methods:

1. Binarize the image by using Otzso's methods.
2. Remove the noise around the boundary.
3. Apply thinning operation.
4. Apply Density based feature extraction.
5. Apply distance-based feature extraction.
6. Normalize both features to lie between 0 and1.
7. Store all the computed features in vector.

C. Script Recognition

HMM classifier is adopted for recognition purpose. HMM is a statistical Markov Model in which the system being modeled is assumed to be a Markov process with hidden state. Table 1 shows the worst letters recognized and their corresponding classification accuracy. The problem of recognizing a handwritten character as a whole can now be

No.	Character	Accuracy (%)	Often mistaken for
1	nga ང	71%	da ད
2	cha ཆ	70%	tsha ཚ
3	ca ཅ	68%	tsa ཙ
4	ja ཇ	75%	dza ཛ
5	pa པ	76%	ba བ
6	ta ཏ	79%	ha ཧ

Table 1. Worst Characters recognized.

Figure 4: Character after preprocessing.

considered as a sequence of decisions in which feature vectors are grouped into smaller decision units and sequentially recognized. The sequence of these decision units represents the unknown word. To solve such a recognition problem, Hidden Markov Models are widely used.

IV. RESULTS AND DISCUSSION

Exactly 70 samples of each characters are prepared for the training proposed and 30 for the testing propose are presented to the Hidden Markov Model classifier, for the dataset, each dataset we have considered each different character, identification of the test data is done using HMM classifier. The results were found to be satisfactory when manually compared. As a whole for Tibetan Handwritten character we obtained the accuracy 86%. Matlab software tool is used for implementation. It was also observed that the feature obtained by density and distance of pixel gives more accurate than the geometric algorithms for Tibetan handwritten characters.

Table 2. Recognition accuracy of Individual characters with two different classification methods.

NO	Character	Classification Method		NO	Character	Classification Method	
		HMM	DCT			HMM	DCT
1		95%	91%	16		93%	95%
2		92%	89%	17		68%	71%
3		91%	90%	18		70%	65%
4		71%	67%	19		75%	73%
5		68%	59%	20		80%	81%
6		70%	71%	21		96%	91%
7		75%	74%	22		97%	93%
8		97%	89%	23		91%	87%
9		79%	74%	24		96%	88%
10		96%	97%	25		97%	91%
11		71%	73%	26		90%	90%
12		89%	83%	27		95%	91%
13		76%	87%	28		96%	89%
14		93%	78%	29		79%	80%
15		76%	74%	30		98%	92%

Table 3. Recognition accuracy of four vowels with two different classification methods.

No	Character	Classification Methods	
		HMM	DCT
1		96%	91%
2		94%	90%
3		91%	87%
4		97%	93%

Table 2 and 3 shows the results of two different classification methods as per 30 different characters and 4 vowels,

From table 2 and 3 we can say that HMM gives better classification result for Tibetan Characters than DCT even DCT gives more accurate value for few characters.

V. CONCLUSION

In this paper, density and distance-based feature extraction system are used to extract the features of Tibetan handwritten characters, and this methodology helps to get a more accurate results than the other methods like geometric based features extraction. HMM has been used for classification. The recognition rate of 86% has been achieved for the Tibetan handwritten character. Experimenting the created dataset with different classifiers to obtain more accuracy will be the future work.

REFERENCES

1. M. R. Kumar, R. Pradeep, B. S. P. Kumar, and P. Babu, “A Simple Text-line segmentation Method for Handwritten Documents,” Int. J. Comput. Appl., no. April, pp. 46–51, 2012.
2. Rachana R. Herekar, “Handwritten Character Recognition Based on Zoning Using Euler Number for English Alphabets and Numerals\n,” IOSR J. Comput. Eng., vol. 16, no. 4, pp. 75–88, 2014.
3. Long-Long Ma and Jian Wu 2016. “Online unconstrained Tibetan character recognition using handwritten statistical recognition” Himalayan Linguistics, Vol. 15(1):1544-7502
4. F. Hedayati, J. Chong, K. Keutzer, and C. Sciences, “Recognition of Tibetan Wood Block Prints with Generalized Hidden Markov and Kernelized Modified Quadratic Distance Function,” 2010.
5. Z. Rowinski and K. Keutzer, “Namsel: An Optical Character Recognition System for Tibetan Text,” Himal. Linguist., vol. 15, no. 1, pp. 10–30, 2016.
6. M. Hangarge, “Gaussian Mixture Model for Handwritten Script Identification,” Int. Conf. Emerg. trends Electr. Commun. Inf. Technol., pp. 64–69, 2012.
7. J. D. Student, “Recognition of Handwritten Script By Using Neural Network,” vol. 2, no. Iii, pp. 1–4, 2014.
8. M. June, C. Mizan, T. Chakraborty, and S. Karmakar, “Available Online at www.ijarcs.info Text Recognition using Image Processing,” vol. 8, no. 5, pp. 765–768, 2017.
9. a. M. Elgammal and M. a. Ismail, “Techniques for language identification for hybrid Arabic-English\ndocument images,” Proc. Sixth Int. Conf. Doc. Anal. Recognit., pp. 1100–1104, 2001.
10. N. Dave and G. H. P. College, “Segmentation Methods for Hand Written Character Recognition,” vol. 8, no. 4, pp. 155–164, 2015.
11. Y. Es-saady, M. Amrouch, A. Rachidi, M. El Yassa, and D. Mammass, “Handwritten Tifinagh Character Recognition Using Baselines Detection Features,” vol. 5, no. 4, pp. 1177–1182, 2014.
12. A. Singh, “Handwritten English Character Recognition using HMM , Baum-Welch and Genetic Algorithm,” vol. 7, no. 4, pp. 1788–1794, 2016.

13. E. Krevat and E. Cuzzillo, “Improving Off-line Handwritten Character Recognition with Hidden Markov Models,” Citeseer, pp. 1–6, 2006.
14. V. Märgner, H. El Abed, M. Pechwitz, V. Märgner, H. El Abed, M. Pechwitz, O. Handwritten, A. Word, V. Märgner, H. El, and A. Mario, “Offline Handwritten Arabic Word Recognition Using HMM - a Character Based Approach without Explicit Segmentation To cite this version : HAL Id : hal-00112048 Offline Handwritten Arabic Word Recognition Using HMM - a Character Based Approach without Explic,” 2006.
15. C. Li and G. Biswas, “Using Hidden Markov Model Based,” Graph. Models, vol. 3, no. 9, pp. 245–256, 1999.
16. A. Choudhary, R. Rishi, and S. Ahlawat, “Off-line Handwritten Character Recognition Using Features Extracted from Binarization Technique,” AASRI Procedia, vol. 4, pp. 306–312, 2013.

AUTHORS PROFILE

Mr. Nyichak Dhondup, MSc. Computer Science, Christ (Deemed to be University), year (2017-2019).
BSc. CMS(Computer Science, Mathematics, Statistics), Christ (Deemed to be University), year (2014-2017).

Ms Deepa V Jose completed her PhD in Computer Science from Christ University. Her areas of interest in research are Internet of Things, Artificial intelligence, Sensor networks, Network Security and Image Processing.