International Journal of Future Generation Communication and Networking
Vol. 12, No. 5, 2019 pp.107-113

Character Segmentation and Identification Methods for Japanese Document Images Using K-Nearest Neighbor Method

K.Sheikdavood, A Nithya
Assistant Professor,
Department of Electronics and Communication Engineering,
M.Kumarasamy College of Engineering, Thalavapalayam, Karur, Tamil Nadu.

Abstract

Now-a-days there are many languages exist in our country. Scope of the Japanese language learning increasing in our country. In general the scripts are affected by their arrangement, style, low print feature and intermixed content like device printed and manuscript. In order to overcome these drawbacks we are using character segmentation and character identification algorithm. Initially the character segmentation algorithm will choose the segmentation line by structural property. Maximum curvature method is used to separate the merged character in a document. Then SVM classifier is used in last step to segment the image. Next the character identification algorithm is used. In this algorithm the geometrical features are calculated. Based on the center pixel character the first and second features are formed. The nearest pixel around the center pixel will help us to calculate the third feature. The character identification algorithm will use the K-NN classifier. The SVM and K-NN classifier is more accurate in segmentation process when compared with other segmentation techniques. The accuracy of this classifier will be 99% when compared with other classifiers.

Index Terms: SVM classifier, K-NN, character segmentation, Maximum curvature method.

I. INTRODUCTION

The computer will analyses the printed or written text by using Automated Optical Character Recognition (OCR).[5] This optical character recognition should recognize the scanned images. This includes photo scanning of the content character-by-character, investigation of the filtered in picture, and afterward interpretation of the character picture into character codes, for example, ASCII, regularly utilized in image processing. [7] This technique is also used in food manufacturing process. In industries the optical character recognition technique is depends on camera quality whereas in document analysis optical character recognition technique is depends on the machine. Uncertain light inside the machine due to shadow, checking point and ink dissemination are a portion of the fundamental issues experienced while examining.[1] There are three processes in optical character recognition technique. They are content localization, character segmentation and character identification. Character identification is an important process in document images for printed content.[10] But the merged content in a multilingual atmosphere it will be difficult task. The problem may occur due to the size and mode of a text. Before identifying the character the document should go to the segmentation process. This segmentation process will split the character from word images.[2] Latin script is similar with English language which has 26 letters, 5 vowels and 21 consonant as like English language shown in fig.1.

Fig.1(a) Latin script

The vowels, consonants and consonant modifiers may also present in Devanagri letters. The Devanagri script contains 13 vowels, 34 consonants and 14 vowel modifiers given in fig.1(b).

ISSN: 2233-7857 IJFGCN

Fig.1 (b) Devanagri Script

Initially the character segmentation algorithm will choose the segmentation path by structural property.[3] Graph distance theory is used to separate the merged character in a document. Then SVM classifier is used in last step to segment the image. Next the character identification algorithm is used. In this algorithm the geometrical features are calculated.[5] Based on the center pixel character the respective features are formed. The nearest pixel around the center pixel will help us to calculate the third feature. [7] The character identification algorithm will use the K-NN classifier. Remaining of this paper will be organized as follows: section II explains the proposed work and next section will discuss the simulation results and conclusion of this paper is given in section IV.

II. PROPOSED WORK

Submit In this section the character segmentation and a character identification algorithm was discussed. SVM algorithm is used to segment the input image. With the help of segmentation process the character is identified by using K-Nearest Neighbour algorithm.[6] The proposed work of this paper will be given in fig.2

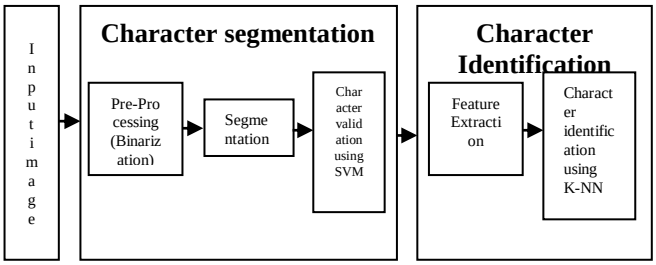

Fig.2 Block Diagram

Initially the document should be scanned and the input image is taken. The character in input image should be identified with better accuracy.[4] For that the given input image will goes under further process like character segmentation and character identification. The input image is given in fig.3

Fig.3 Input Image

A. *Character Segmentation*

The input image will be processed and moves for segmentation process. In the segmentation process the image

will be binarized first and it will be segmented. First the input image is binarized and the vertical projection profile of an image should be considered with pixel intensity of $f_{th}\ (m,n)$ by using the following formula,

$$h_n = sum_{m=1}^{x} f_{th}\ (m,n) \quad (1)$$

Where h_n is a no.of.vertical projection path. The noise and fake projection paths may also generated. In order to remove the noise and fake lines Gaussian low pass filter is used. After removing the noise the image line is changed to smoothen projection profile. The smoothen line whose range from 0 to 2 was retained.[3] These projection lines will be coordinated with input image. At that time over-segmentation may occur that can be overcome by using subsequent conditions.

(a) In the event that distinction of two projection lines is not exactly or level with to limit, hold the correct one and force out the left one.

(b) If contrasts of two bulge lines are more prominent than this limit, a normal of two bulge lines is determined.

(c) Yet, if issue of over-division continues and at whatever point contrast of two bulge lines is more prominent than limit, hold the left one. [7]

In the above step the joined character was segmented in exact way.[9] If the characters are merged then it is difficult to segment. For this the maximum curvature method is used. The characters which is equivalent breadth and division picture element are high power pixels.[5] Presently, beginning in descending bearing and in every emphasis, present pixel turns into the middle picture element of a seeking window 3×3, which is appeared in Fig. 4.[12] As inquiry is in the descending course, just next column picture element are considered to diminish the excess and to make seeking instrument quick and basic.[13]

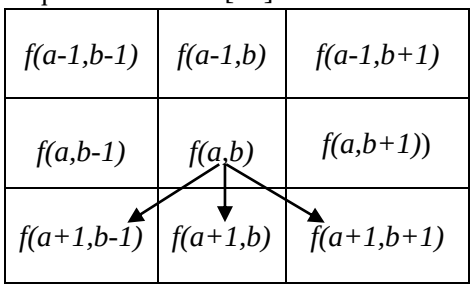

Fig.4 Searching Method For Finding Segmentation Pixel

The input image detects the character region and the binarization process will takes place by maximum curvature method.[11] Then the binarized character was extracted by repeated line tracking method shown in fig.5

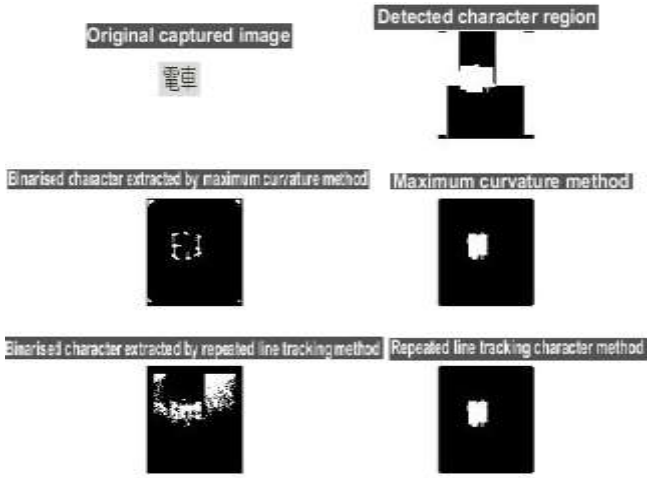

Fig.5 Binarized Character

The over segmentation may occur during binarization process. Even the conditions may apply to reduce this problem the over segmentation will exist. In order to overcome that SVM classifier is used.[14] Linear Kernel SVM algorithm is used with high dimensional data. By using this classifier the input image was segmented. There are two sections in segmented process.[8] They are correct and incorrect section in training database. From the

input image the character is extracted and segmented and goes to dilation process. The segmented image is feed to trained SVM for dilation process.[6] The output of the segmented character is accurate by using SVM algorithm by removing the over-segmentation problem. Fig.6 explains the segmentation process with dilation.

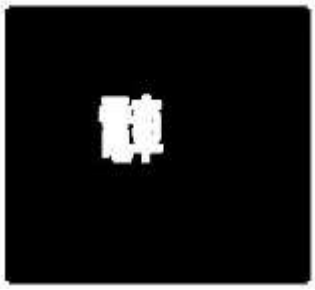

Fig.6 Dilated Image

B. *Character Identification*

The character identification section should perform the feature extraction process and uses the K-NN algorithm for character recognition. The characters should be classified for feature extraction.[10] For character identification K-NN algorithm is used. For this basic leadership step, k-Nearest Neighbor is utilized in the projected character identification work where highlights of information character is contrasted and highlights of preparing tests to figure comparable k neighbors. [6] Exactness of k-NN relies on two elements, (i) remove work utilized and (ii) estimation of k. Here, three separations, for example Euclidean, city square what's more, cosine are utilized to figure separate between an information include vector and each preparation highlight vector. Euclidean separate is given as,

$$D(a,b) = \sqrt{\sum_{i=1}^{N}(a_i - b_i)^2} \quad (2)$$

City square distance is calculated by

$$C(a,b) = \sum_{i=1}^{N}|a_i - b_i| \quad (3)$$

Cosine distance is calculated by

$$C_{cos}(a,b) = \frac{\sum_{i=1}^{N} a_i b_i}{\sqrt{\sum_{i=1}^{N} a_i^2}\sqrt{\sum_{i=1}^{N} b_i^2}} \quad (4)$$

Here, N denotes measurement of quality vectors and distances (D, C, and C_{cos}) between input vector a and bi are considered using (2-4). At this instantaneous, k nearest distances are take into account and number of times input vector belonging to each class are calculated.[13] Output of classifier is the class that occurs very frequently shown in fig.7.

Fig.7 Processed Binary Image

III. SIMULATION RESULTS

A. *Performance of character segmentation process*

Initially the input image is analyzed and pre-processing step is carried out. In pre-processing step the binarization takes place. The input image is consisting of noise. The noise was removed by using Gaussian low

pass filter.[6] The maximum curvature method is used for binarization process. The binarized character is extracted by frequent line tracking technique. The maximum curvature method and repeated line tracking method acting an important role in segmentation process. During binarized character extraction the over-segmentation problem may occur. To overcome that support vector machine algorithm is used.[2] The SVM algorithm is used for segmentation to improve the accuracy. The experimental result of segmentation process in shown in fig.8

Fig.8 Result of Character Segmentation

B. ***Performance of Character Identification process***

The output of character segmentation is next moves for feature extraction process. Based on the center pixel character the first and second features are formed.[4] The nearest pixel around the center pixel will help us to calculate the third feature. The character identification algorithm will use the K-NN classifier. The distance between the pixels is calculated by using Euclidean distance, city distance and cosine distance. The identification rate is measured by

$$IR = \frac{No.of\ correctly\ classified\ samples}{Total\ no.of\ samples} \times 100 \quad (5)$$

From the feature extraction the character is identified by using K-NN algorithm. The extracted character is given to the K-NN algorithm. This algorithm will improve the accuracy and minimizes the time consumption.[9] At last the character will be identified by using character identification algorithm with better accuracy shown in fig.9 and fig.10

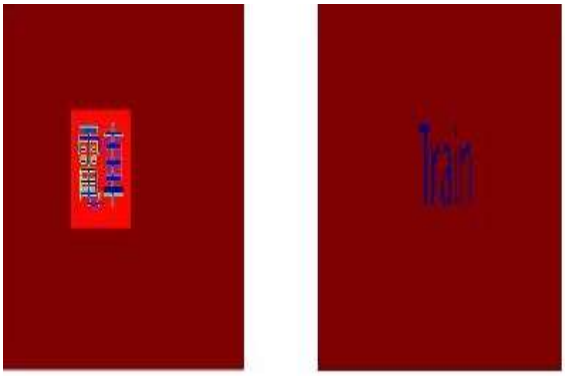

Fig.9 Result of Character Identification Process

	Recognition Schemes for Indian Document Identification
sensitivity	95.9986
specificity	99
Accuracy	99.0007
Processing_Time	0.1237

Fig.10 Recognition Methods For Indian Document

IV.CONCLUSION

For Japanse hand written and printed text documents the character segmentation and character identification algorithm is used. The documents may contain different scripts. The scripts may have the problem with low quality images and merged texts. For this problem we are going for segmentation process. The character of input image is segmented by using the support vector machine algorithm. This SVM algorithm is used to reduce the over-segmentation problem in segmentation process. Then feature extraction process is carried out by next step. Then the character identification process will be executed by using K-NN algorithm. Character in the input image is identified by K-NN algorithm. The character identification and segmentation algorithm used in this proposed system is more efficient when compared with other database. The accuracy of the character identified is about 99%.

REFERENCES

1. Manikandan M, Andrews N V, Kavitha V (2018) Investigation on Micro Calification Of Breast Cancer From Mammogram Image Sequence. International Journal of Pure and Applied Mathematics 118(20): Pages 645-649
2. K. S. Dash, N. B. Puhan, and G. Panda, ‘‘Unconstrained handwritten digit recognition using perceptual shape primitives,’’ in Pattern Analysis and Applications. London, U.K.: Springer, Nov. 2017.
3. S.Palanivel Rajan, C.Vivek "Analysis and design of microstrip patch antenna for radar communication", Journal of Electrical Engineering and Technology, Online ISSN No.: 2093-7423, Print ISSN No.: 1975-0102, Impact Factor–0.597, 2019.
4. Keerthi S, Dhivya S (2017) Comparison of RVM and SVM Classifier Performance in Analysing the Tuberculosis in Chest X Ray. International Journal of Control theory and Applications 10(36): Pages 269-276.
5. Rajan S P, Vivek C, Paranthaman M (2016) Feasibility Analysis of Portable Electroencephalography Based Abnormal Fatigue Detection and Tele-Surveillance System. International Journal of Computer Science and Information Security 14(8): Pages711-722
6. P. Sahare and S. B. Dhok, ‘‘Review of text extraction algorithms for scene- text and document images, Apr. 2016
7. Khanduja, N. Nain, and S. Panwar, ‘‘A hybrid feature extraction algo- rithm for devanagari script,’’ ACM Trans. Asian Low-Resour. Lang. Inf. Process.,Jan. 2016.
8. K. Verma and R. K. Sharma, ‘‘Comparison of HMM-and SVM-based stroke classifiers for Gurmukhi script,’’ Neural Comput. Appl., Dec. 2016.
9. Rajan, S., & Paranthaman, M. (2019). Characterization of compact and efficient patch antenna with single inset feeding technique for wireless applications. *Journal of Applied Research and Technology*, *17*(4).
10. Paranthaman, M., and S. Palanivel Rajan. "Design of Triple C shaped Slot Antenna for Implantable Gadgets." *Current Trends In Biomedical Communication And Tele–Medicine* (2018): 40. DOI: 10.21786/bbrc/11.2/6
11. N. R. Soora and P. S. Deshpande, ‘‘Novel geometrical shape feature extraction techniques for multilingual character recognition,’’ Oct. 2016.

12. S.Palanivel Rajan, M.Paranthaman, Dr.C.Vivek, (2016) "Design and Enhancement of Wideband Reconfigurability using Two E-Shaped Patch Antenna", Asian Journal of Research in Social Sciences and Humanities, ISSN : 2249-7315, Vol.6, Issue 9, pp. 317-327

13. M Paranthaman, G.Shanmugavadivel "Design of Frequency Reconfigurable E-Shaped Patch Antenna for Cognitive Radio" International Journal of Applied Engineering Research, ISSN 0973-4562 Vol. 10 No.20 (2015) pp.16546-16548

14. S. Tian et al., ''Multilingual scene character recognition with co- occurrence of histogram of oriented gradients,'' Mar. 2016.

15. S.Palanivel Rajan, et.al., "Performance Evaluation of Mobile Phone Radiation Minimization through Characteristic Impedance Measurement for Health-Care Applications", IEEE Digital Library Xplore, ISBN : 978-1-4673-2047-4, IEEE Catalog Number: CFP1221T-CDR**,** 2012**.**

16. S.Palanivel Rajan, et.al., **"**Experimental Explorations on EOG Signal Processing for Real Time Applications in LabVIEW**",** IEEE Digital Library Xplore, ISBN : 978-1-4673-2047-4, IEEE Catalog Number: CFP1221T-CDR**,** 2012**.**

17. K. Sheikdavood, M. Ponni Bala," Similarity Identification of an Image using Various Filtering Techniques," International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue-6S3, April 2019

18. N. R. Soora and P. S. Deshpande, ''Robust feature extraction technique for license plate characters recognition,'', Jan. 2015.

19. S.Palanivel Rajan, V.Kavitha, "Diagnosis of Cardiovascular Diseases using Retinal Images through Vessel Segmentation Graph", Current Medical Imaging Reviews, Online ISSN: 1875-6603, ISSN: 1573-4056, Vol. : 13, Issue :4, DOI : 10.2174/1573405613666170111153207, 2017.

20. S.Palanivel Rajan, "Review and Investigations on Future Research Directions of Mobile Based Tele care System for Cardiac Surveillance", Journal of Applied Research and Technology, Vol.13, Issue 4, pp.454-460, 2015.

21. N. Das, R. Sarkar, S. Basu, P. K. Saha, M. Kundu, and M. Nasipuri, ''Hand- written Bangla character recognition using a soft computing paradigm embedded in two pass approach,'', Jun. 2015.

22. Rajan S. P, Paranthaman M. Novel Method for the Segregation of Heart Sounds from Lung Sounds to Extrapolate the Breathing Syndrome. Biosc.Biotech.Res.Comm. 2019;12(4).

23. M Paranthaman, S Vijayprasath, S Palanivel Rajan "Design of a Frequency Tunable Patch Antenna using HFSS"International Journal of Advanced Research Trends in Engineering and Technology, Vol.3, Issue 7 (2016) pp.69-72

24. J. R. Prasad and U. Kulkarni, ''Gujrati character recognition using weighted k-NN and Mean X 2 distance measure,'' Int. J. Mach. Learn. Cybern., Feb. 2015.

25. Paranthaman, M., and S. Palanivel Rajan. "Design of Triple C shaped Slot Antenna for Implantable Gadgets." *Current Trends In Biomedical Communication And Tele–Medicine* (2018): 40. DOI: 10.21786/bbrc/11.2/6

26. K. C. Santosh and L. Wendling, ''Character recognition based on non- linear multi-projection profiles measure,'' Frontiers Comput. Sci., Oct. 2015.

27. K.Sheikdavood, S.Palanivel Rajan, "Analysis of Ovarian Diseases Using Ultrasound Images", Journal of advances in chemistry , Vol. 12, Issue 10, pp. 4449-4454, 2016.

28. S.Palanivel Rajan, K.Sheik Davood, "Performance Evaluation on Automatic Follicles Detection in the Ovary", International Journal of Applied Engineering Research, Vol.10, Issue 55, pp.1-5, 2015.