IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-5, Issue-9 E-ISSN: 2347-2693

Extraction of Tamil Characters from a Handwritten Document using Connected Component Labeling

D. Rajalakshmi[1*], S.K. Jayanthi[2]

[1*]Dept. of Computer Science, Vellalar College for Women, Bharathiar University, Coimbatore, India
[2]Dept. of Computer Science, Vellalar College for Women, Bharathiar University, Coimbatore, India

[*]*Corresponding Author: rajalakshmi.d@gmail.com, Tel.: +91-9994552514*

Available online at: www.ijcseonline.org

Received: 04/Aug/2017, Revised: 18/Aug/2017, Accepted: 09Sep/2017, Published: 30/Sep/2017

Abstract – Writer identification is a challenging task for the reason that it requires textural features and structural features. Textural features like grey-level co-occurrence matrices, Gabor filters can be extracted from entire page or a block of text. The structural features like slant and skew, character height, stroke width, frequency of loops or blobs etc. also characterize the handwriting style. Before extracting character level features it is a prerequisite to segment the document image into characters. This paper proposes a connected component oriented approach to segment an image of handwritten Tamil document into individual characters. The features extracted from these characters then can be used for writer identification.

Keywords: Writer Identification, Handwritten documents, Segmentation, Connected Component, Tamil Script.

I. INTRODUCTION

The identification of a person on the basis of digitized form of handwriting is a useful biometric modality with forensic and historic document analysis and continues to be an exemplary study area within the research field of behavioral biometrics. Handwriting of a person has unique features and can be used for identification. The recent advancements in image processing, data mining, pattern recognition and machine learning have proved that it is possible to automate writer identification.

In order to characterize a writing style, writer specific features are essential. Features are quantitative measurements that can be obtained from a handwriting sample. These features can be obtained from the whole document or from a block of text or from a character [1]. Connected components, enclosed regions, lower and upper contours and fractal features are features at character level to represent a handwriting style [2].

Combining textural and structural features yields increased writer identification rate [3]. Segmentation of a handwritten document into components is essential for feature extraction process. To obtain character level features, it is necessary to segment the document image into characters.

Historically, recognition started with position-based, pixel-oriented segmentation methods, and the current approach is pattern orientation. Pattern- oriented methods have made it possible to segment written characters and even touching handwritten characters more flexibly [4].

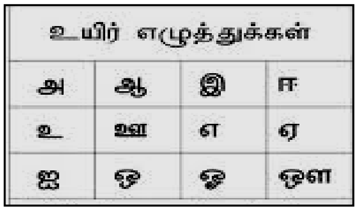
The ten official Indic scripts - Devanagari, Tamil, Gurmukhi, Kannada, Telugu, Guajarati, Bengali, Oriya, Malayalam and Urdu - most of which have not seen much targeted research in human language technologies, regardless of the large number of users. Tamil, the native language of Tamil Nadu state in India has several million speakers across the world and considered as an official language in countries such as Sri Lanka, Malaysia and Singapore.

In this paper, a method based on connected components is proposed to segment a Tamil handwritten document image into characters. The paper is organized as follows. In Section 2, the characteristics of Tamil script are presented. In Section 3 the related work is discussed. In Section 4, the proposed work is explained. Section 5 exhibits the results and Section 6 concludes the work.

II. CHRACTERISTICS OF TAMIL SCRIPT

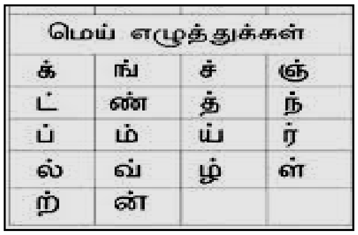
The Tamil script is written from left to right. It has 12 vowels (“uyireluttu” or "soul-letters"), 18 consonants (“meyyeḻuttu” or "body-letters") and one special character (“ayutha eluttu”). The complete script, consists of the 31 letters in their independent form and an additional 216 combinations (“uyirmeyyeḻuttu” or “soul-body-letters") for a total of 247. The combination letters are formed by adding a vowel marker to the consonant. Some vowels require the basic shape of the consonant to be altered in a way that is specific to that vowel. Others are written by adding a vowel-specific suffix to theconsonant, yet others a prefix, and still other vowels require adding both a prefix and a

suffix to the consonant.



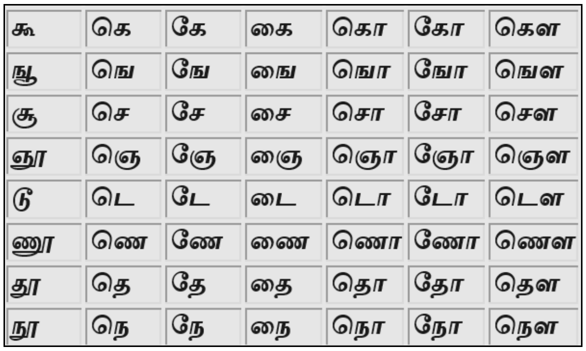
உயிர் எழுத்துக்கள்			
அ	ஆ	இ	ஈ
உ	ஊ	எ	ஏ
ஐ	ஒ	ஓ	ஔ

Figure 1: a) Tamil Vowels



மெய் எழுத்துக்கள்			
க்	ங்	ச்	ஞ்
ட்	ண்	த்	ந்
ப்	ம்	ய்	ர்
ல்	வ்	ழ்	ள்
ற்	ன்		

Figure 1: b) Tamil Consonants



கூ	கெ	கே	கை	கொ	கோ	கௌ
ஙூ	ஙெ	ஙே	ஙை	ஙொ	ஙோ	ஙௌ
சூ	செ	சே	சை	சொ	சோ	சௌ
ஞூ	ஞெ	ஞே	ஞை	ஞொ	ஞோ	ஞௌ
டூ	டெ	டே	டை	டொ	டோ	டௌ
ணூ	ணெ	ணே	ணை	ணொ	ணோ	ணௌ
தூ	தெ	தே	தை	தொ	தோ	தௌ
நூ	நெ	நே	நை	நொ	நோ	நௌ

Figure 1: c) Sample Combinational Characters

III. RELATED WORK

Segmentation of handwritten text document into lines, words and characters is one of the most important and challenging tasks in handwriting analysis. Segmentation of individual characters is a straightforward process when characters are well spaced. This problem becomes much more difficult when characters are touching or overlapping.

Richard G. Casey and Eric Lecolinet presented a review on character segmentation strategies. The character segmentation methods are classified into i) Classical approach ii) Rule based iii) Hybrid iv) Holistic approach [5]. The classical approach partition the image into sub images which is called "dissection". The second class of methods avoids dissection, and segments the image either by classification of prespecified windows, or by classification of subsets of spatial features collected from the image. The third one is a hybrid of the first two methods. Holistic approaches recognize entire character strings as units.

Bansal and Sinha proposed a two pass algorithm for the segmentation of Devanagari composite characters into their constituent symbols [6]. The algorithm is based on structural properties of the script. In the first pass, words are segmented into characters. Statistical information about the height and width of each divided component is used to hypothesize whether a component is composite. In the second pass, the hypothesized composite components are further segmented. The algorithm is designed to segment a pair of touching characters.

Munish Kumar, M. K. Jindal, and R. K. Sharma applied vertical projection profile and water reservoir based technique for identification and segmentation of touching characters in handwritten Gurumukhi words. Touching characters are segmented based on reservoir base area points [7]. Saba et al. Provided a survey on methods for touching character segmentation. They divide the touching character segmentation techniques into two classes that perform explicit or implicit character segmentation [8].

Ye Xiangyun, Mohamed Cheriet, and Ching Y. Suen proposed a stroke model to depict the local features of character objects as double-edges in a predefined size. This model enables to detect thin connected components selectively, while ignoring relatively large backgrounds that appear complex [9]. Sridevi and Subashini presented a method in which connected components algorithm is combined with nearest neighborhood algorithm to segment the characters of ancient Tamil scripts [10].

Vassilis Papavassiliou et al. presented two approaches to extract text lines and words from handwritten document. The line segmentation algorithm is based on locating the optimal succession of text and gap areas within vertical zones using Viterbi algorithm. A text-line separator drawing technique is applied and then finally the connected components are assigned to text lines [11].

Dharmapryia C. Bandara et al. proposed an algorithm that guarantees the segmentation of individual handwritten character skeletons into meaningful segments, avoiding the problems of over-segmentation and under-segmentation [12]. Mamatha and Srikantamurthy proposed a segmentation scheme for segmenting handwritten Kannada scripts into lines, words and characters using morphological operations and projection profiles [13]. Siddhartha Banerjee et al. proposed a segmentation algorithm based on region growing for extraction of characters in a bank cheque [14]

He Lifeng et al. presented the importance of connected component labeling process. Connected component labeling is indispensable for distinguishing different objects in a binary image, and prerequisite for image analysis [15].

The study of related work reveals that segmentation of individual characters is a direct process when characters are well spaced. As Tamil Script is non-cursive in nature, the individual characters in a word are isolated. Spacing between characters can be used for segmentation. If the characters are treated as objects, the connected component labeling method can be applied for segmenting the characters.

IV. PROPOSED WORK

In this section, segmentation of unconstrained handwritten Tamil script into characters is proposed. The method consists of two phases, prepocessing and extraction of individual characters. There is no standard database available for Tamil handwritten samples and a database is created by own. The writers were asked to write text lines in A4 size pages. No constraints were forced regarding the content and use of pen. Writers chosen were Undergraduate students of a College. The scripts were collected and scanned using flatbed scanner at a resolution of 300 dpi and stored as jpg format images.

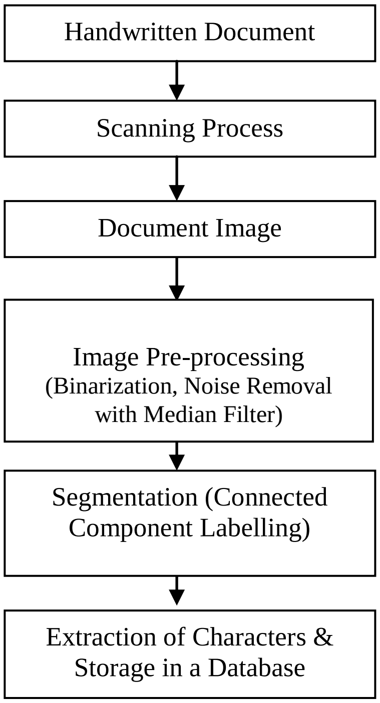

Handwritten Document
Scanning Process
Document Image
Image Pre-processing
(Binarization, Noise Removal with Median Filter)
Segmentation (Connected Component Labelling)
Extraction of Characters & Storage in a Database

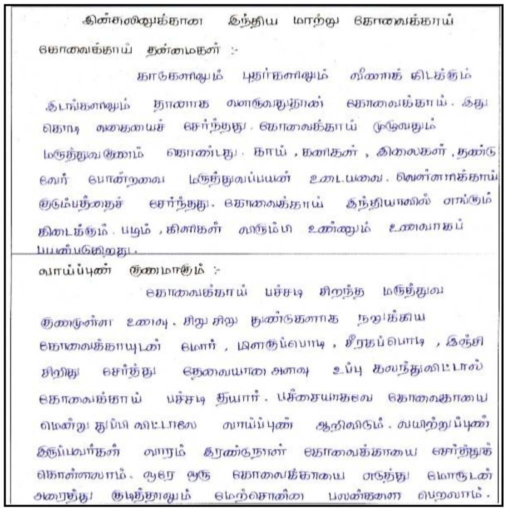
Figure 2. Steps in Extraction Process.

Our character extraction methodology is shown in Figure. 2 which depicts various phases in segmentation and character extraction.

A. Preprocessing

Preprocessing techniques enhance the image features and support for better feature extraction. In our attempt, preprocessing techniques include the tasks such as binarization and noise removal. Image preprocessing refers to the operations on images at the lowest level of abstraction. The aim of preprocessing is to improve the image data that suppresses undesired distortions and enhances some image features that are relevant for further processing and analysis.

The binarization process converts the grayscale image into binary form which facilitates several tasks such as document analysis, script recognition and character recognition. Connected component labeling technique requires the image format as binary. The grayscale image is converted to binary form using Otsu's method [16].



இன்சுலினுக்கான இந்திய மாற்று கோவைக்காய்

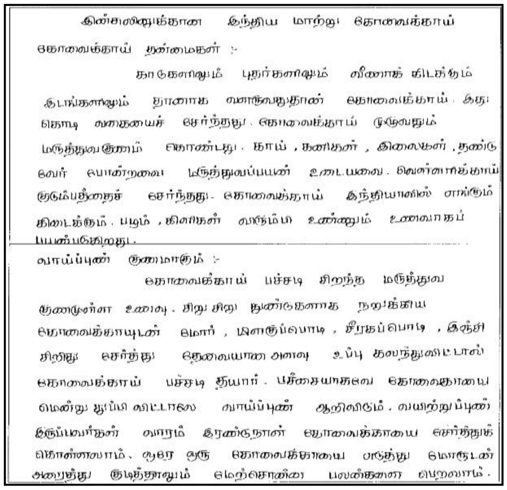
கோவைக்காய் தன்மைகள் :-

காடுகளிலும் புதர்களிலும் வீணாக் கிடக்கும் இடங்களிலும் தானாக வளருவதுதான் கோவைக்காய். இது கொடி வகையைச் சேர்ந்தது. கோவைக்காய் முழுவதும் மருத்துவ குணம் கொண்டது. காய், கனிகள், இலைகள், தண்டு வேர் போன்றவை மருத்துவப்பயன் உடையவை. வெள்ளரிக்காய் குடும்பத்தைச் சேர்ந்தது. கோவைக்காய் இந்தியாவில் எங்கும் கிடைக்கும். பழம், கிளிகள் விரும்பி உண்ணும் உணவாகப் பயன்படுகிறது.

வாய்ப்புண் குணமாகும் :-

கோவைக்காய் பச்சடி சிறந்த மருத்துவ குணமுள்ள உணவு. சிறு சிறு துண்டுகளாக நறுக்கிய கோவைக்காயுடன் மோர், மிளகுப்பொடி, சீரகப்பொடி, இஞ்சி சிறிது சேர்த்து தேவையான அளவு உப்பு கலந்துவிட்டால் கோவைக்காய் பச்சடி தயார். பச்சையாகவே கோவைக்காயை மென்று துப்பி விட்டாலே வாய்ப்புண் ஆறிவிடும். வயிற்றுப்புண் இருப்பவர்கள் வாரம் இரண்டுநாள் கோவைக்காயை சேர்த்துக் கொள்ளலாம். ஒரே ஒரு கோவைக்காயை எடுத்து மோருடன் அரைத்து குடித்தாலும் மேல்சொன்ன பலன்களை பெறலாம்.

Fıgure: 3 (a) Original document image



இன்சுலினுக்கான இந்திய மாற்று கோவைக்காய்

கோவைக்காய் தன்மைகள் :-

காடுகளிலும் புதர்களிலும் வீணாக் கிடக்கும் இடங்களிலும் தானாக வளருவதுதான் கோவைக்காய். இது கொடி வகையைச் சேர்ந்தது. கோவைக்காய் முழுவதும் மருத்துவ குணம் கொண்டது. காய், கனிகள், இலைகள், தண்டு வேர் போன்றவை மருத்துவப்பயன் உடையவை. வெள்ளரிக்காய் குடும்பத்தைச் சேர்ந்தது. கோவைக்காய் இந்தியாவில் எங்கும் கிடைக்கும். பழம், கிளிகள் விரும்பி உண்ணும் உணவாகப் பயன்படுகிறது.

வாய்ப்புண் குணமாகும் :-

கோவைக்காய் பச்சடி சிறந்த மருத்துவ குணமுள்ள உணவு. சிறு சிறு துண்டுகளாக நறுக்கிய கோவைக்காயுடன் மோர், மிளகுப்பொடி, சீரகப்பொடி, இஞ்சி சிறிது சேர்த்து தேவையான அளவு உப்பு கலந்துவிட்டால் கோவைக்காய் பச்சடி தயார். பச்சையாகவே கோவைக்காயை மென்று துப்பி விட்டாலே வாய்ப்புண் ஆறிவிடும். வயிற்றுப்புண் இருப்பவர்கள் வாரம் இரண்டுநாள் கோவைக்காயை சேர்த்துக் கொள்ளலாம். ஒரே ஒரு கோவைக்காயை எடுத்து மோருடன் அரைத்து குடித்தாலும் மேல்சொன்ன பலன்களை பெறலாம்.

Fıgure: 3 (b) Result after Preprocessing

Noise can appear in a document image during the conversion process and also caused by dirt on the document. Noise can be removed by simple filters like median or morphological operators. Median filter is applied to remove noise. Figure 3 shows the document image and its binary form after removal of noise.

B. Connected Component Labeling

Segmentation is a process of distinguishing characters of a hand written document, a crucial step as it extracts the meaningful regions for analysis. In Document Image Processing, four commonly used segmentation algorithms are Connected Component labeling, X-Y tree decomposition, Run–length smearing and Hough transform.

Connected component is a technique which assigns to each component of the binary image a distinct label. The background pixel is treated as 0 and foreground pixel is taken as 1. The labels are usually natural numbers starting from one to the total number of connected components in the image [17] [18]. A connected component in a binary image is a set of pixels that form a connected group.

The connected component labeling process scans the image from left–to–right and top-to-bottom. On the first line containing black pixels, a unique label is assigned to each contiguous run of black pixels. For each black pixel of the next and succeeding lines, the neighbouring pixels of the previous line and the pixel to the left are examined. If any of the neighbouring pixels has labeled, the same label is assigned to the current pixel; otherwise the next unused label is used. The sequential algorithm for connected component labeling requires two passes over the image.

```
Step 1: Scan the image left to right, top to bottom.
Step 2: If the pixel is foreground, then
    2.1 If only one of its upper or left neighbours has a
        label, then copy the label.
    2.2 If both have same label, then copy the label.
    2.3 If both have different labels, then copy the upper
        label and enter the label in the equivalence table.
    2.4 Otherwise assign a new label to this pixel and
        enter this label in the equivalence table.
Step 3: If there are no more pixels to consider, go to
    Step 2.
Step 4: Find the lowest label for each equivalent set in
    the equivalence table.
Step 5: Scan the image and replace each label by the
    lowest label in its equivalent set.
```

The procedure to extract individual components in the image is given below

```
Step 1: Find the number of connected components
    in the Image.
Step 2: For each connected component
    2.1 Find minimum and maximum row and
        column values (Top-left and Right-bottom).
    2.2 Store the component as separate image
        using above mentioned values.
Step 3: Repeat Step2 until all components are
    extracted.
```

V. RESULTS

Text lines from different handwriting samples from the database [19] were considered for experiments of the character segmentation. The procedure for connected component labeling is implemented using MATLAB tool. The characters extracted from a handwritten image sample are shown in Figure 5.

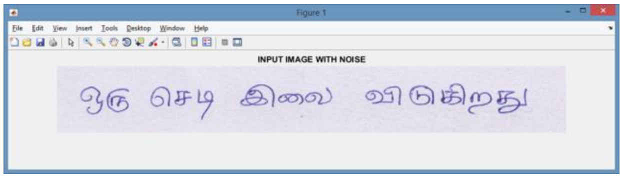


Figure: 5 (a) Input Image

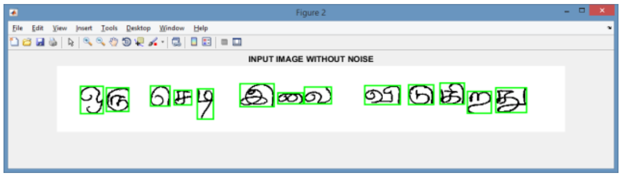


Figure 2
File Edit View Insert Tools Desktop Window Help
INPUT IMAGE WITHOUT NOISE
ஒரு செடி இலை விடுகிறது

Figure: 5 (b) Connected Components

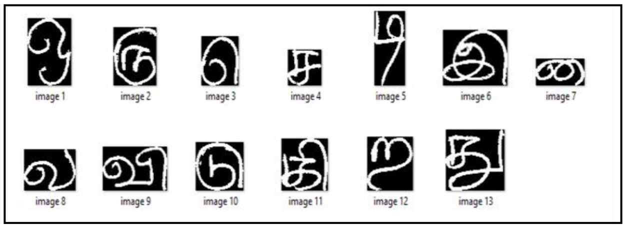


Figure: 5 (c) Extracted Character components

If the sample image contains combinational characters, this method segments it into sub characters. This is shown in Figure 6. In case of character recognition problems, this may be a limitation. If physical features of characters are considered, the sub components with least dimensions (dot components) may be discarded and other sub components (prefix or suffix components) can be treated as individual characters.

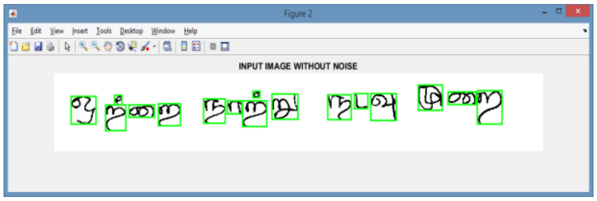


Figure 2
File Edit View Insert Tools Desktop Window Help
INPUT IMAGE WITHOUT NOISE

Figure 6: a) Input sample - Combinational characters

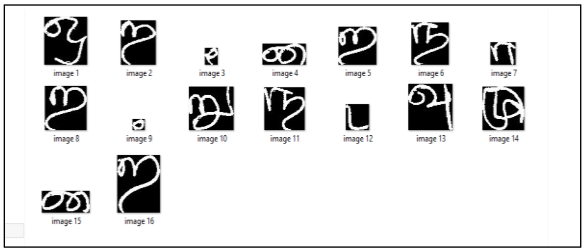


Figure 6: b) Segmented character components
(Combinational characters subdivided into sub components)

VI. CONCLUSION

The issue of character extraction from handwritten document images is still one of the challenging topics in the field of document image analysis. This paper addresses the problem of segmenting characters from Tamil handwritten documents. Individual character components of document image is required for extracting features to serve the purpose of writer identification. The proposed method, segments the characters by connected component labeling approach. This method provides robust results and the characters extracted successfully are adequate for writer identification.

REFERENCES

[1] Srihari, Sargur N., Sung-Hyuk Cha, Hina Arora, and Sangjik Lee. "*Individuality of handwriting.*" Journal of Forensic Science, Vol.47, No. 4,pp.1-17,2002.

[2] Hertel, Caroline, and Horst Bunke. "*A set of novel features for writer identification.*" International Conference on Audio-and Video-Based Biometric Person Authentication. Springer, Berlin, pp. 679-687, 2003.

[3] Bulacu, Marius, and Lambert Schomaker. "*Combining multiple features for text-independent writer identification and verification.*" In Proc. of Tenth International Workshop on Frontiers in Handwriting Recognition, La Baule, pp. 281–286, 2006.

[4] H. Fujisawa, Y.Nakano and K.Kurino, "Segmentation methods for character recognition from *segmentation to document structure analysis*", Proceedings of the IEEE, Vol.80, No.7, pp. 1079-1092, 1992.

[5] Richard G. Casey and Eric Lecolinet, "*A Survey of Methods and Strategies in Character Segmentation*", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.18, No. 7, pp. 690-706, 1996.

[6] Bansal V, Sinha R M K., "*Segmentation of Touching and Fused Devanagari Characters*." Pattern Recognition.,Vol.35, No.4, pp. 875-893,2002.

[7] Munish Kumar, M. K. Jindal, and R. K. Sharma. "Segmentation of isolated and touching characters in offline handwritten Gurmukhi script recognition." International Journal of Information Technology and Computer Science (IJITCS),Vol.6,No.2,p.58, 2014.

[8] Saba, Tanzila, Ghazali Sulong, and Amjad Rehman. *"A survey on methods and strategies on touched characters segmentation*." International Journal of Research and Reviews in Computer Science,Vol. 1,No.2, pp. 103-114, 2010.

[9] Ye Xiangyun, Mohamed Cheriet, and Ching Y. Suen. "Stroke-model-based character extraction from gray-level document images." IEEE Transactions on Image Processing, Vol.10,No.8, pp.1152-1161,2001.

[10] Sridevi, N., and P. Subashini, "*Segmentation of Text Lines and Characters in Ancient Tamil Script Documents using Computational Intelligence Techniques."* International Journal of Computer Applications,Vol. 52, No.14,pp.7-12, 2012.

[11] Vassilis Papavassiliou, Themos Stafylakis, Vassilis Katsouros, and George Carayannis, *"Handwritten document image segmentation into text lines and words*". Pattern Recognition, vol. 43, pp. 369 – 377, 2010.

[12] Dharmapryia C. Bandara,Vasile Palade, and Ruskan Batuwita, *"A Customizable Fuzzy System for Offline Handwritten Character Recognition,"* International Journal on Artificial Intelligence Tools, vol. 20, no. 3, pp. 425–455, 2011.

[13]Mamatha, H. R., and K. Srikantamurthy. *"Morphological Operations and Projection Profiles based Segmentation of Handwritten Kannada Document."* International Journal of Applied Information Systems (IJAIS), Vol. 4, No.5, pp.13-19,2012.

[14] Siddhartha Banerjee,Bibek Ranjan Ghosh,Arka Kundu ,"*Handwritten Character Recognition from Bank Cheque*",International Journal of Computer Sciences and Engineering ,Vol.4,No.1, pp.99-104,2016.

[15] He, L., Ren, X., Gao, Q., Zhao, X., Yao, B. and Chao, Y. *"The connected-component labeling problem: A review of state-of-the-art algorithms*". Pattern Recognition,Vol. 70,pp. 25-43, 2017.

[16] N. Otsu, "*A threshold selection method from grey level histogram*". IEEE transactions on Systems, Man, and Cybernetics, Vol. 9, pp. 62-66, 1979.

[17] Rafael C.Gonzalez , Richard E.Woods, "*Digital Image Processing*",Third Edition, Dorling Kindersley India Pvt. Ltd., India, pp. 645-647,2009.

[18] Gurpreet Kaur and Jaskaranjit Kaur, "*A Comparative Study of Image Demosaicing*", International Journal of Computer Sciences and Engineering, Vol.3, Issue.7, pp.98-102, 2015.

[19] D.Rajalakshmi,S.K.Jayanthi,*"Collection of Offline Tamil Handwriting Samples and Database Creation"*, International Journal of Advanced Research in Computer and Communication Engineering,Vol. 5, Issue 8, pp. 196-199, 2016.

Authors Profile

Dr. S.K.Jayanthi received the M.Sc., M.Phil.,PGDCA.,Ph.D in Computer Science from Bharathiar University in 1987,1988,1996 and 2007 respectively. She is currently working as an Associate Professor and Head , Department of Computer Science in Vellalar College for Women,Erode,Tamilnadu,India. She secured District First Rank in SSLC under Backward Community. Her research interest includes Image procesing, Pattern recognition and Fuzzy Systems. She has guided 34 M.Phil scholars and 2 Ph.D scholars. Currently 2 M.Phil scholars and 7 Ph.D scholars are pursuing their degree under her supervision.She is a member of ISTE, IEEE and Life member of Indian Science Congress. She has published 40 research papers in reputed journals and she has presented 80 papers in International/National conferences.

D.Rajalakshmi received Bachelor of Science from Bharathiar University Coimbatore and Masters from Bharathidasan University. She is a part time Doctoral research scholar in Department of Computer Science, Vellalar College for women, Erode, Tamilnadu,India and currently working as Assistant Professor in Department of Computer Science, Vidyasagar College of Arts and Science, Udumalpet, Tamilnadu, India. She is qualified with SET (Statelevel Eligiblility Test) for Lectureship. She has published one research paper in International Journal and pressented two papers in International conferences.Her main research work focuses Document Image Processing, Machine Learning and Data Mining.