RESEARCH ARTICLE OPEN ACCESS

A Comparative Study of Optical Character Recognition for Printed and Handwritten Tamil Text

Persia A
Assistant Professor Vidhya Sagar Women's College, Chengalpattu, Tamilnadu, India

ABSTRACT
This paper compares the methodologies for recognizing printed Tamil text document with handwritten Tamil text. Each of the preprocessing phases uses various techniques. But both the techniques are used to extract the features of Tamil characters. In printed Tamil documents, the extracted features are passed to a Support Vector Machine (SVM) where the characters are classified by Supervised Learning Algorithm and then these are mapped onto Unicode for recognition. In the case of handwritten Tamil text, Support Vector Machine (SVM) is used for recognizing the characters. This paper analysis the performance of these two methods and presents a report based on the analysis.

Keywords: *Tamil Character Recognition, Support Vector Machine (SVM), Preprocessing, Unicode, Feature Extraction.*

I. INTRODUCTION

OCR for Tamil character is software, which converts printed form of Tamil characters and the image form of Tamil characters into system compatible word processing document in Tamil. It saves the time effort of developing a digital copy of any document. This process can turn any document into several formats such as Microsoft Word, Excel, HTML, PDF or Rich Text Formats. These documents are editable and allows user to modify the content. Character Recognition can be divided into printed text and handwritten text [5]. Ancient literary to latest works can be converted into system readable.

So the people who are all interested in Tamil, able to get the traditional books through Internet. The remainder of this paper is organized as follows:

Section II presents the background review and related work, which are important for the understanding of this paper. Section III describes the methods and algorithms that are used in the OCR for Tamil printed text and handwritten text. In section IV various experimental results are discussed. Here the OCR recognized different handwritten characters written by different users and in the case of printed Tamil text, character features are extracted and classification also done in this section. Finally, the comparisons between these two are presented using graphs.

II. RELATED WORK

"Optical Character Recognition for printed Tamil text using Unicode" [1] describes the efficient method for recognizing printed Tamil characters that are converted into software translated Unicode Tamil text which is based on Support Vector Machine (SVM). Here the characters are classified using Supervised Learning Algorithm.

"A Complete OCR for printed Tamil Text" [3] tells about the recognition of printed Tamil text in a different manner. A multi-rate –signal-processing based algorithm is used to achieve distortion. The images of the words are subjected to morphological closing followed by connected-component based segmentation to separate out the individual symbols. A three-level, tree-level structured classifier Tamil script is designed.

"Handwritten Tamil Character Recognition using SVM" [4], this author discussed the same techniques of R. Seethalakshmi et al. [1]

"A Novel SVM-based handwritten Tamil Character Recognition System" [2] proposed a summary for recognizing offline handwritten Tamil characters using Support Vector Machine (SVM). Various preprocessing operations are performed on the digitized image to enhance the quality of the image.

III. OCR FOR TAMIL

OCR is the process of converting printed Tamil text into software translated Unicode Tamil text.

A. Printed Tamil Text

The block diagram of OCR consists of various stages as shown in Fig 1. Those are scanning phase, preprocessing, segmentation, feature extraction, classification (SVM, rule based and ANN), Unicode Mapping and recognition and output verification.

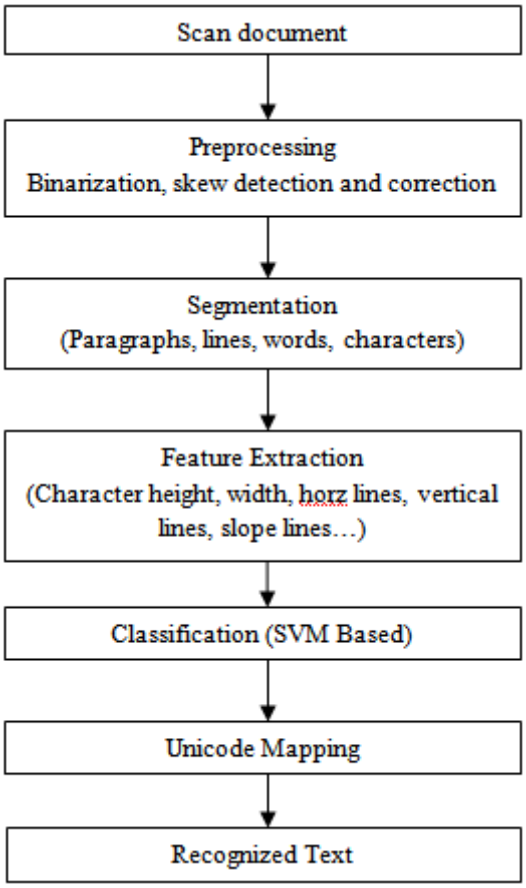


Fig 1. Block Diagram of OCR [1]

Scanning and preprocessing the document

A printed document is placed over the standard scanner for scanning. It scans the document and sent to a program that saves it in TIR, JPG, or GIF format. In the preprocessing stage the image file is checked for skewing. The skewed image is corrected using simple rotation technique in the appropriate direction [3]. Here, the noise is eliminated from the image and binarized.

Segmentation

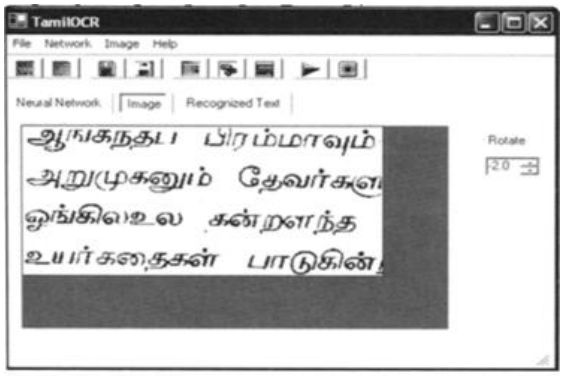
After the preprocessing phase the image is segmented using an algorithm which decomposes the scanned text into paragraph using vertical histogram and lines into words using horizontal histograms and words into character image glyphs using horizontal histogram. Each image glyph consists of 32×32 pixels. Histogram means bar chart representing a frequency distribution; heights of the bars represent observed frequencies.

(a)

Fig 2(a). Histogram for skewed and skew corrected images [1]

(b)

Fig 2(b). Skewed image [1]



(c)

Fig 2 (c) Skew corrected and preprocessed image [1]

The above Fig 2(a) gives the histogram for skewed and skew corrected image, Fig 2(b) gives the skewed image and Fig 2(c) gives the skew corrected and preprocessed image [1].

Feature Extraction

Here, the features of each image glyph are extracted. The various features are taken into consideration for classification.

First a character glyph is define by the following attributes: Height of the character, Width of the character, Numbers of horizontal lines present-short and long, Numbers of vertical lines present-short and long, Numbers of circles present, Numbers of horizontally oriented arcs, Numbers of vertically oriented arcs, Centroid of the image, Position of the various features, Pixels in the various regions [1].

The various feature extraction algorithms are follows.

- Horizontal line detection
- Vertical line detection
- Slope lines detection
- Detection of circles and arcs

Classification

The extracted features of characters are analyzed using the set of rules.

- A typical rule based classifier
  If ((Numbers of short horizontal lines==0) and (No of long horizontal lines==1) and (Numbers of short vertical lines==0) and (Numbers of long vertical line==1) and (Numbers of circles==1) and (Numbers of Horizontally oriented Arcs==1) and (Numbers of Vertically Oriented Arcs==1))
  Then the character is **m**
- Back propagation based Classifier
  Here a Back propagation based Artificial Neural Network is chosen for classification because of its simplicity and ease of implementation [8]. The architecture consists of three layers: Input, hidden and output.
- Support Vector Machine (SVM) based

The Support Vector Machine (SVM) is tested to recognize the printed Tamil character.

- Kernel Functions

There are a number of kernels that can be used in Support Vector Machine. These include linear, polynomial, radial basis function (RBF) and sigmoid.

- Unicode Mapping

After classification the characters are recognized and a mapping table is created in which the Unicode for the corresponding characters are mapped.

- Character Recognition

The scanned image is compared with the recognition from the mapping table from which corresponding Unicode are accessed.

B. Handwritten Tamil Text

Support Vector Machine (SVM) can also be used for recognizing handwritten Tamil text. Handwritten Tamil character recognition based on Support Vector Machine (SVM). There are three tasks are involved in character recognition. Those are Preprocessing Feature Extraction and Classification.

Preprocessing

This phase consist of Thresholding, Skeletonization, Line segmentation, Character recognition, Normalization.

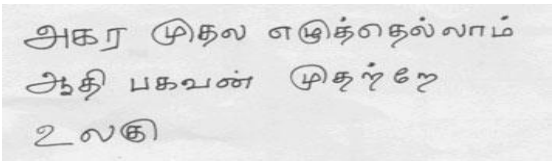


Fig 3. Input Image [2]

Thresholding is used to extract the foreground from the background. Here Otsu's method of histogram-based global Thresholding algorithm is used [6]. Skeletonization is the process of peeling off a pattern without affecting the general shape. Thinning algorithms are used to convert offline handwriting to online data. A number of thinning algorithms are available. Here, Hilditch algorithm is used for Skeletonization [7]. It is a parallel-sequential algorithm which checks all the pixels at one pass at the same time and decisions are made. Line separation is an operation to separate individual lines of text from the document image. The horizontal histogram is calculated for the binary image and histogram is used to segment the lines. Character segmentation is an operation which decomposes an image of sequence of characters into sub images of individual characters. The vertical histogram profile is calculated for each segmented line. The task of normalization is to converting the random sized image into standard sized image [2]. All the segmented character images are normalized into a common height and width (32×32 pixels) using bilinear interpolation technique. Fig 4 gives the normalized image.

Fig 4. Normalized Image [2]

Feature Extraction and Recognition

It is the problem of extracting the information from the preprocessed data which is most relevant for classification purposes. Various recognition methods are available for handwriting recognition. In this paper SVM is used for recognizing handwritten Tamil characters.

IV. RESULTS AND DISCUSSIONS

In this section we discuss about the experimental results for printed and handwritten Tamil text recognition based on Support Vector Machine (SVM) method. The performance of SVM is compared with the typical rule based and Back propagation Network (BPN).

A. Printed Tamil Text

Here the Optical Character Recognition (OCR) is implemented in Microsoft.Net. Various experimental results are discussed and various comparative studies were also taken. These are Back propagation based classifier, Typical rule based and Support Vector Machine (SVM). From the Table 1the following observations are made. Typical rule based and BPN when a single font is taken, it gives 100% accuracy where as multiple fonts doesn't give the 100% efficiency. This accuracy is reduced to some extent. But in the case of Support Vector Machine (SVM), it gives good performance. Five font types like Arial Unicode MS, Anjal (Nalinam), Amudham, Elango, Shree_tam fonts are considered for recognition [1].

Fig 5. Output of Segmentation [1]

TABLE 1 Comparison of classifiers [1]

Types of classifier	Fonts	Efficiency
Typical rule based	1	100%
Typical rule based	2 or more	80%
BPN	1	100%
BPN	2 or more	50%
SVM	Single font	100%
SVM	Multi font	100%

B. Handwritten Tamil text

The system is trained with 35,441 characters belonging to 106 different Tamil characters written by 117 different users. SVM is trained using the set of data and recognition accuracy is calculated for the test data. In this experiment 4×4 zone is chosen which produces high recognition accuracy. When the zone's size is small, it captures more detailed pixel variations. There is high dissimilarity because of varying nature of handwriting which is depicted in the Fig 6. Large sized zones failed to capture the essential parts of characters, which make distinct from others. The best results were produced by 4×4 pixel zone [2].

Class	Character	No. of Samples	Recognition Accuracy %	Class	Character	No. of Samples	Recognition Accuracy %
1	அ	182	81.32	18	ன	182	75.27
2	ஆ	182	78.02	19	த	181	83.98
3	இ	183	62.84	20	ந	166	75.30
4	ஈ	177	83.05	21	ப	179	95.53
5	உ	179	88.83	22	ம	182	90.11
6	ஊ	180	71.11	23	ய	182	90.66
7	எ	173	82.63	24	ர	182	78.57
8	ஏ	181	69.06	25	ல	181	91.16
9	ஐ	178	88.76	26	வ	178	90.45
10	ஒ	172	70.35	27	ள	171	66.08
11	ஓ	173	64.16	28	ழ	178	77.53
12	ஃ	182	98.90	29	ற	177	91.53
13	க	180	91.11	30	ன	175	65.14
14	ங	167	87.43	31	ர	178	82.58
15	ச	177	84.18	32	ஷ	181	89.50
16	ஞ	175	66.86	33	ஜ	175	89.71
17	ட	183	97.81	34	ஹ	176	87.5
					Overall	6048	82.04

Fig 6. Recognition accuracy for 34 Tamil texts [2]

C. Comparison between Printed and Handwritten Tamil character

Based on the above discussion, the printed and handwritten Tamil characters are compared here. The algorithms which were used by these papers are taken into consideration. The accuracy or efficiency of OCR depends on the algorithm we deployed. When we talk about the accuracy, printed Tamil documents only gives better performance than handwritten while using Support Vector Machine (SVM). But in the case of handwritten, it fails to do this. So, we can use the fusion algorithms to improve the efficiency of recognition of handwritten Tamil documents [5]. Here we give a comparison between printed Tamil documents and handwritten Tamil documents using OCR which is based on SVM in a graphical manner Fig 7.

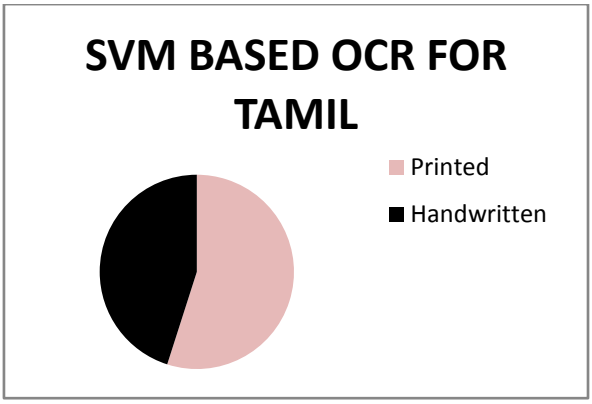


Fig 7. Recognition using SVM

V. CONCLUSION

This paper deals with Support Vector Machine (SVM) for recognizing printed and handwritten Tamil documents. From the experiment and results, Support Vector Machine (SVM) is suitable for both printed and handwritten. But, if fusion algorithms are used in printed Tamil documents, it gives maximum accuracy. When we analysis these two methods, the SVM is not at all suitable for handwritten Tamil recognition. So in future, we

can also go for fusion of algorithms in handwritten Tamil characters to improve the efficiency.

REFERENCES

[1] Seethalakshmi. R, Sreeranjani T.R., Balachandar T., "Optical Character Recognition for printed Tamil Text Using Unicode", Journal of Zhejiang University SCIENCE, pp. 1297-1305, 2005.
[2] N. Shanthi and K. Duraisamy, "A novel SBM-based handwritten Tamil character recognition system", Springer – Verlag London Limited, pp. 173-180, 2009.
[3] A.G. Ramakrishnan and Kaushik Mahata, "A complete OCR for Printed Tamil Text", pp. 151-156.
[4] Dr. J. Venkatesh and C. Sureshkumar, "Handwritten Tamil Character Recognition Using SVM", International Journal of Computer and Network Security, pp. 29-33, 2009.
[5] R. Jagadeesh Kannan and R. Prabhakar, "A Comparative Study of Optical Character Recognition for Tamil Script", European Journal of Scientific Research, pp. 570-582, 2009.
[6] R. Jagadeesh Kannan and R. Prabhakar, "Off-Line Cursive Handwritten Tamil Character Recognition", Issue 6, Volume 4, pp. 351-360, June 2008.
[7] Anbumani Subramanian and Bhadri Gubendran, "Optical Character Recognition for Printed Tamil Characters", Dec 2000.
[8] K.H. Aparna and V.S. Chakravarthy, "A Complete OCR system development of Tamil Magazine Documents", Tamil Internet, pp.45-51, 2003.

International Journal of Engineering Research and Applications (IJERA) is UGC approved Journal with Sl. No. 4525, Journal no. 47088. Indexed in Cross Ref, Index Copernicus (ICV 80.82), NASA, Ads, Researcher Id Thomson Reuters, DOAJ.

Persia A. " A Comparative Study of Optical Character Recognition for Printed and Handwritten Tamil Text" International Journal of Engineering Research and Applications (IJERA) 7.8 (2017): 56-60.