Research Paper

Volume : 3 | Issue : 4 | Special Apr Issue 2014 • ISSN No 2277 - 8179

Technology & Innovation

KEYWORDS : OCR, Tamil Characters

A Survey on Character Recognitions in Tamil Scripts using OCR

T. S. Suganya | Research Scholar, Sathyabama University, Chennai, India

Dr. S. Murugavalli | Professor and Head, CSE Department, Panimalar Engineering College, Chennai, India

ABSTRACT

since a computer is a rather fast machine, it's possible to check an incoming glyph against all possible characters, and classify the input as the best match. In order to do this the program must be trained to recognize all possible characters with emergence the need for the development of a high performance OCR engine has become essential. Difficulties are faced by modern readers in interpreting an ancient script. This paper analyses the various approaches and challenges in recognitions of Tamil characters and issues related to recognitions are discussed.

I. Introduction

Historical documents are usually posing much degradation, due to weather conditions, preservation and handling methods etc.. It is very important to preserve the historical documents which reveal the information about the civilized part. One of the most important applications in image processing and pattern Recognition is Optical Character Recognition. OCR is a very well-studied problem in the vast area of pattern recognition. The first commercial OCR systems began to appear in the early 1950s and soon they were being used by the US postal service to sort mail.

A script may be used by only one language or may be shared by many languages, sometimes with slight variations from one language to other. There are lots of difficult things which can be solved through image processing technique, separating each character, recognizing character fonts and written styles used in different centuries. Many researchers try to apply many techniques for breaking through the complex problem of Tamil character recognition.

II. Tamil Language

Tamil is one of the oldest languages in the world with rich literature. Recognition of any handwritten characters with respect to any language is difficult. Tamil the native language of a southern state in India has several million speakers across the world and is an official language in several countries. Tamil script recognition differs from other language scripts in a few significant ways. It is stole or lost over time, so we have to process the existing documents.

III. Preprocessing

Preprocessing involves noise removal, skew detection and correction and Binarization of the gray-valued digital images.

A. Noise Removal

Noise is defined as any degradation in the image due to external disturbance. Quality of handwritten documents depends on various factors including quality paper, aging of documents, and quality of pen, color of ink etc.

B. Skewing:

The Source Images often suffer from Noise of different sizes, so the source image is preprocessed for noise removal and the resultant image is checked for Skewing as there are possibilities of image getting skewed with either left or right orientation.

C. Binarization

It is the process of converting a gray scale image into Binary image by Thresholding. During the Thresholding process, individual pixels in an image are marked as "object" if their value is greater than some threshold value and as "background" pixels if their value is less than threshold value.

D. Segmentation

Segmentation means the separation of character from non-character in the given source image. It subdivides an image into its constituent objects. The level of detail to which the subdivision is carried depends on the problem being solved. The process includes separating the preprocessed image into lines, words and characters. The accuracy of the OCR system depends on the segmentation.

IV. Feature Extraction

The attributes of a character which make it distinct from other characters are called the features. The process of obtaining them from individual characters is called Feature Extraction.

A. Correlation-based features

It involves distance measures which are computed with a point-by-point analysis of the input characters. Further, such features are not even invariant to transformations.

B. Transform-based features

It employs a transform the axes of representation in order to emphasize certain attributes of the input characters. Some of the transforms which have been employed for this purpose are the Fourier transfer, the Karhumem-Loeve transform, the Wavelet transform and the Harr transform.

C. Statistical features

This is derived from statistical distribution of point and/or characters include zoning, moments, cumulates and characteristic loci. They provide high speed and low computational complexity and are invariant to changes in font face.

D. Geometrical features

This is also called Topological features. It is extracted from the shape of the input characters. Some typical features are the strokes, lines and relative positions of strokes. These features are highly tolerant to most types of distortions.

E. Feature Selection

Here it involves the choice of the right features for the given problem is done based on the structural and statistical properties of all the input patterns. Once the right features representative of the characters are extracted. The neighborhood approach for classification identifies the neighbors of the current feature point in the feature space. Shiromone et al. [4] have proposed an encoded character string dictionary for recognition of Tamil characters. This principle of theirs is taken by Chandrasekaran et al. [5,6] for recognition of constrained hand printed Tamil characters and later for multifont Tamil character recognitions. Identification at the character level, Recognition of ancient characters from inscriptions is difficult. Rajakumar [19] concentrates on the century identification of ancient Tamil characters and converting them into current century's form using MATLAB. Panayiotis et al [27] pair wise matching of all realizations of an alphabet symbol on two inscriptions.

The detection of cracks within the stones surface Hubert Mara et al [22] will enable the separated processing of partially preserved characters. Hang et al. (2012)[10] has discussed the misty, foggy, or hazy weather conditions lead to image color distortion and reduce the resolution. Sridevi et al.[34] PSO in which an optimal threshold is calculated for segmenting the text

lines with nearest neighborhood algorithm to segment the characters. S.K.Thilagavathy et al[35] Dilation and filtration used to find character.

LuShijian et al.[26] report a pair of document vectorization technique and their applications to the identification of scripts and Document vectorization is accomplished by using character. Utpal Garain et al [28] (ICA) based image enhancement technique is presented to improve the accuracy for machine reading of camera-based images as a 3-class problem. Nizar et al[30] characterize Arabic and Latin ancient document images from Fractal dimension method is used to discriminate between these two scripts. Gangamma et al [33] bilateral filter which averaging without smoothing the edges. Hubert Mara et al [22] anisotropic filtering using standard computer hardware to preserve the characters of eroded ancient Chinese characters carved into stone.

V. Classification

The extracted features are employed to make a decision on the class to which the test pattern belongs. The most widely used classifiers are the Neural Network (multi-layer perceptron), Kernel methods including Support Vector Machines (SVM), K nearest neighbours, Gaussian mixture model, Gaussian, naïve bayes, decision trees and RBF classifiers, classification algorithms include Linear classifiers (Fisher's Linear Discriminant, Logistic regression, Naïve Bayes classifier, Perceptron), Support Vector Machines (Least Squares Support Vector Machines), Quadratic Classifiers, Kernel estimation (k-nearest neighbor), Boosting, Decision trees (Random forests), Neural networks, Gene Expression Programming, Bayesian networks, Hidden Markov models, Learning vector quantization.

VI. Table: Comprehensive Study

Researchers	Methods Features	Classifiers	Scope of application	Best recogniti on reported
K.H. Aparna et al[18]	*Structure or shape based*	*FSA*	*Handwritten*	*86.4%*
N. Dhamayanthi et al[17]	*Neural Networks*	*Back Propagation Network*	*Handwritten*	*90%*
Rajakumar et al[2][32][19][9]	*Contour-let transform*	*Clustering mechanism*	*Recognition of individual character*	*NA*
	Slant angle estimation & correction	*Lower base line & Upper base line detection*		*95.74%*
	Fourier-Wavelet Coefficients	*Nearest Neighbour*		*98%*
	SIFT	*K-means*		*84%*
Niranjan Joshi et al[16]	*Preprocessed x-y co-ordinates, quantized slope values & dominant point co-ordinates*	*Template based elastic matching algorithm*	*Handwritten*	*95.9%*

Seethalakshmi et al[15]	*SVM*	*Supervised Learning algorithm*	*Unicode fonts*	*NA*
Dr. N.Sengottaiyan et al[3]	*Physical, Topological, Mathematical or Statistical*	*Hybrid*	*Handwritten*	*97%*
Dr.B.P. Malikarjunaswamy et al[1]	*Graph pyramid*	*Bottom-up*	*Graph representation of a character*	*NA*
Jagadeesh Kumar et al[13,29]	*Time Domain Features Frequency Domain Features Dehooking algorithm*	*Hidden Markov Model Nearest neighbor Classifier*	*Handwritten*	*98%*
Suresh Kumar C et al [14]	*SVM*	*Neural Basis Function Network, Hybrid Nero Fuzzy System, RCS Algorithm*	*Handwritten*	*97%*
Nirasefathima et al[23,31]	*Standard Structural Features Antonacopoulos& Parkers shape tracing*	*Structural DBN Classifier*	*Distorted Tamil Character*	*NA*
R.M.Suresh et al[24]	*Distances of the pattern from the frame in 16 different directions*	*Fuzzy Approach*	*Handwritten numerals and Tamil characters*	*94%*

*NA: Not Available – Recognition results not given in terms of Numerical values.

VII. Outlook

The paper gives an overview of the approaches that were used in the ongoing research in Tamil character recognition field. Though each of the above discussed methods has their own superiorities and drawbacks, the recognition accuracy rates are reported to be above 85%. The researcher is to guarantee the accuracy and security of information extraction.

VIII. Conclusion

This paper has presented a related work on OCR Techniques. Most of the research works already reported is explained in deep.

The following challenges can be further explode in my research work:

- Text should be clearly extracted from its background to obtain a good recognition result for the characters.
- Variations in writing style of characters in different centuries.
- Difficulties in the lack of database and so on during character recognition processes.
- Recognitions of the characters which would assist historians and archeologist to know the cultural heritage of the civilization so as to enable further exploration.

REFERENCE

[1] Dr. B.P.Mallikarjunaswamy and Karunakara K, " Graph Based Approach for Background Elimination and Segmentation of the Image", Research Journal of Computer Systems Engineering- An International Journal Vol 02, Issue 02, June 2011. | [2] Dr. Subbiah Bharathi V and Rajakumar S, " 7th Century Ancient Tamil Character Recognition from Temple wall Inscriptions", Indian Journal of Computer Science and EngineeringVol. 3 Oct-Nov 2012. | [3] Dr.N.Sengottaiyan and Dr. C.Sureshkumar, " Handwritten South Indian Lanhuage Recognition Using Artifical Neural Network", International Journel of Advancd Research In TechnologyVol. 1 Issue 1, 2011, 87-90. | [4] V.K. Govindan and A.P.Shivaprasad, "Character recognition – a review", Pattern Recognition, vol. 23, no. 7, pp. 671-683, 1990. | [5] M. Chandrasekaran, R. Chabdrasekran, and G. Siromeoney, " Context dependent recognition of hand-printed Tamil characters", in Prjoc. Int. conf. on Systems, Man and Cybernatics (India), col. 2, pp. 786-790, 1984. | [6] M.Chandrasekaran, R.Chandrasekran and G.Siromoney, "Computer recognition of Tamil, Malayalam and Devanagari characters", Journal of the IETE(India), vol. 30, pp. 150-154, 1984. | [7] Peeta Basa Pati and A.G.Ramakrishnan, " OCR in India Scripts: A Survey" | [8] Jomy John, Pramod K.V. Kannan Balakrishnan , " Handwritten Character Recognition of South Indian Scripts: A Review", National Conference on Indian Language Computing, Kochi, Feb. 19-20 2011. | [9] Dr. Subbiah Bharathi V and Rajakumar S, "An Off Line Ancient tamil Script Rcognition from Temple Wall I", Indian Journal of Computer Science and EngineeringVol. 3 Oct-Nov 2012. | [10] Yong-Qin Zhang, Yu Ding, Jin-Sheng Xiao, Jiaying Liy and Zongming Guol, " Visibility enhancement Using image filtering approach", EURASIP Journal on Advances in Signal Processing 2012. | [11] Tiji M Jose and Amitabh Wahi , " Recognition of Tamil Handwritten Characters using Daubechies Wavelet Transforms and Fed-forward Back Progapation Network", IJCA, vol. 64, No. 8 pp. 0975-8887, Feb 2013. | [12] D. Ghosh, T.Dube, A.P.Shivaprasad, "Script Recognition – A Review", IEEE Transactions on Pattern Analysis and Machine Intelligence vol.XX, No. 2009 | [13] Jagadeesh Kumar R, Prabhakar R and Suresh R.M, "Off-line Cursive Handwritten Tamil Characters Recognition", International Conference o Security Technology pp 159-164, 2008. | [14] Suresh Kumar C and Ravichandran T, "Handwritten Tamil Character Recogniltion using RCS algorithm", Int. J. of Computer Applications, (0975-8887) vo;. 8-No.8. October 2010. | [15] Seethalakshmi .R, SreeRanjani T.R, Balachandar.T, " Optical Character Recognition for Printed Tamil Text using Unicode", Journal of Zhejiang University Science6A(11):1297-1305 2005. | [16] Niranjan Joshi, G.Sita and A.G.Ramakrishnan, " Cpmparison of Elastic Matching Algorithms for Online Tamil Handwritten Character Recognition", Proceedings of the 9th Int'l Workshop on Froniers in Handwriting Recognition (IWFHR-9 2004). | [17] N.Dhamayanthi, P.Thangavel, " Handwritten Tamil Character Recogntion Using Neural Network", Digitized by Viruba. | [18] K.H.Aparna, Vidhya Subramanian, M.Kasirajan, G.Vijay Prakash, V.S.Chakravarthy, "Online Handwriting Recognition for Tamil", Proceedings of the 9th Int'l Workshop on Frontiers in Handwriting Recognition (IWFHR-9 2004). | [19] S.RajKumar, Dr. V.Subbiah Barathi, "Century Identification and Recognition of Ancient Tamil Character Recognition" Int'l Jour. Of Computr Applications (0975-8887) vol. 26-No. 4, July 2011. | [20] S.K.B.Sangeetha , Dr. V. Vijayachamundeeswari, " Tamil OCR-A Survey", Int'l Con. On Computing and Control Engineering (ICCCE2012),12 &13 April 2012. | [21] Dr. C.P. Sumathi, S.Karpagavalli, " Techniques and Methodologies for Recognition of Tamil Typewritten and Handwritten Characters" A Survey Int'l Jour.of Computer Science & Eng. Survey (IJCSES) vol. 3 No. 6 Dec. 2012. | [22] Hubert Mara, Jan Hering and Susanne Kromker, " GPU based Optical Character Transcription for Ancient Inscription Recognition". | [23] Nirase Fathima Abubacker, Indra Gandhi raman, " An Approach for Structural Feature Extraction for Distorted Tamil Character Recognition", Int'l Jour. Of Computer Applications(0975-8887) vol.22-No.4,May 2011. | [24] R.M.Suresh, L.Ganesan, "Recognition of Printed and Handwritten Tamil Characters Using Fuzzy Approach", Proceedings of the 6th Int'l Con.on Computational Intelligence and Multimedia Applications (ICCIMA) 2005. | [25] P.Banumathi, Dr. G.M.Nasira, "Handwritten Tamil Character Recognition using Artificial Neural Networks", Int.l Conf. on Process Automation, Control and Computing PACC), pp. 1-5, 2011. | [26] Lu Shijian, Chew Lim Tan, " Script and Language Identification in Noisy and Degraded Document Images", IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 30, No. 1, Jan. 2008. | [27] Panayiotis rousopoulos, Michail Panagpoulos, Constantin Papaodysseus, Fivi Panopoulou, for Ancient Inscriptions Writer Identification", IEEE Transaction 978-1-4577-0274-7/11 2011. | [28] Utpal Garain , Atishay Jain, Anjan Maity, Bhabatosh Chanda, " Machine Reading of Camera-Held Loe Quality Text Images: an ICA-Based Image Enhancement Approach for Improving OCR accuracy", IEEE Transaction 978-4244-2175-6 2008. | [29] Ishwarya M.V, R. Jagadeesh Kannan, " An Improved Online Tamil Character Recognition Using Neural Networks" IEEE Transaction 978-0-7695-0/10 2010. | [30] Nizar ZAGHDENI, Remy MULLOT, Adel ALIMI, " Characterization of ancient document images composed by Arabic and Latin scripts", IEEE Transaction 978-1-4577-0314-0/11 2011. | [31] Nirase Fathima Abubacker, Raman Indra Gandhi, " An Extended Method for Recognition of Broken Type written Characters Special Reference to Tamil Script", IEEE Transaction 978-1-61284-931-7/11 2011. | [32] Rajakumar S, Dr. Subbiah Bharathi.V," Ancient t Tamil Script Recognition from Stone Inscriptions using Slant Removal Method",Int.'l Con. On Electrical, Electronics and Biomedical Engineering (ICEEBE'2012) Penang (Malaysia) May 19-20, 2012. | [33] B.Gangamma, Srikanta murthy, Arun vikas singh,"Restoration of Degraded Historical Document Image", Journal of Emerging Trends in Computing and Information Science", vol.3, No. 5, May 2012. | [34] N.Sridevi, P.Subashini, "Segmentation of Text Lines and Characters in Ancient Tamil Script Documents using Computational Intelligence Techniques", Int.'l Journal of Computer Applications, vol.52-N0.14 August 2012. | [35] S.K.Thilagavathy, Dr.R.Indra Gandhi, "Recognition of Distorted Character using Edge Detection Algorithm", Int.'l Journal of Innovative Research in Computer and Communication Engineering vol.1, Issue 4 June 2013. | [36] Indu Sreedevi, Rishi Pandey, N.Jayanthi,Geetanjali Bhola, Santanu Chaudhury, "NGFICA Based Digitization of Historic Inscription Images" ISRN Signal Processing vol.2013. | [37] N.Sridevi, P.Subashini, "Optimized Framework for Classification 11th Century Handwritten Ancient Tamil Scripts using Computational Intelligence Technique" The Int.'l Jor. Of Computer Science & Applications vol.2 No. 2 April 2013. |