ISSN: 2321-788X UGC Approval No: 43960 Impact Factor: 2.114

TAMIL CHARACTER RECOGNITION USING PYTESERACT AND NLTK

Article Particulars

Received: 08.6.2017 Accepted: 25.7.2017 Published: 28.7.2017

Mrs. L. SURIYA KALA
Research Scholar, Department of Computer Science,
Mother Teresa Women's University, Kodaikanal,
Tamil Nadu, India

Abstract

India is a multilingual multiscript nation with in excess of 18 languages and 10 distinctive significant contents. Insufficient research progress in the direction of recognition of transcribed characters of these Indian contents has been finished. Tamil, an official just as well known content of the southern piece of India, Singapore, Malaysia, and Sri Lanka has an enormous character set which incorporates many compound characters. A way to deal with gather utilize full data from an assortment of pictures, in 2014, as indicated by Mary Meeker's yearly Internet Trends report, individuals transferred a normal of 1.8 billion advanced pictures each and every day. That is 657 billion photographs for each year. These pictures can be gathered, put away, handled and broke down for utilizing full information. In this paper proposed to Tamil character recognition utilizing Pyteseract and NLTK process.

Index Terms: Pyteseract, NLTK, Machine Learning, Artifical Intelligence.

1. Introduction

Disconnected recognition of written by hand characters has been examined well in the writing similarly as Latin and a couple of different contents of the created countries are concerned. Overviews of related works are found. Be that as it may, there has not been a lot of progress towards recognition of written by hand characters of Indian contents. Then again, such recognition issue for an Indian content is distinctive in nature in view of the size of its letter set and the likenesses between various characters of an Indian letter set. Additionally, dissimilar to in English content, the letters in order of an Indian content has countless compound characters framed by both vowel-consonant and consonant-consonant blends. Subsequently, the issue of written by hand character recognition of an Indian content needs more consideration. There are bunches of strategies proposed however nobody module to process a wide range of

information having a place with a wide scope of language, in this diary, we are proposing a solitary module which can be the most utilize full approach ever to change over the natural and uncategorized information into utilization full data. An unrefined image of an epigraph contains bothersome pictures or stamps, upheaval introduced and message engraved with a lot of slants.

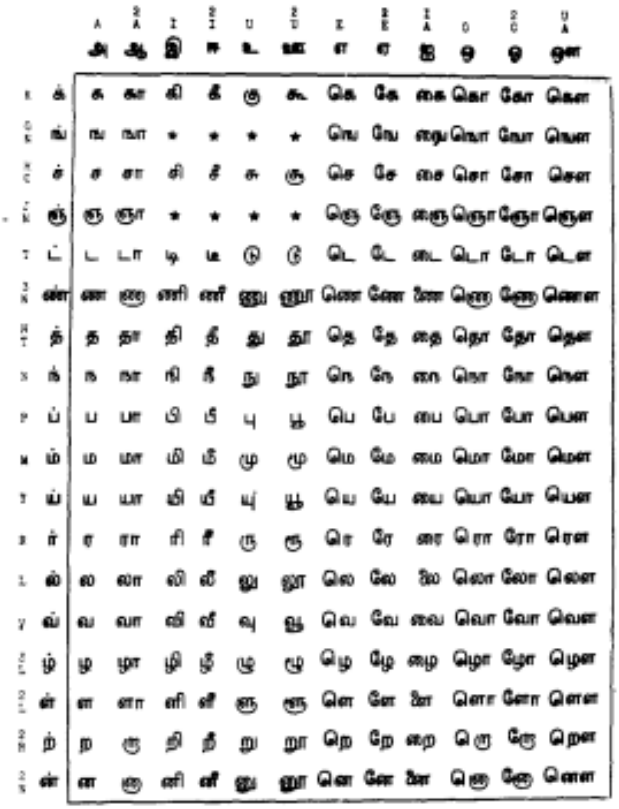

Figure 1 Tamil Alphabet Sets

The isolating among characters besides between the lines and the slant could tangle the path toward deciphering the contents. Some contacting lines and moreover characters jumble the system of division which is a commitment for the affirmation methodology in the later stages. Along these lines, the information file image of epigraphs is to be preprocessed for the departure of disturbance, slant recognition and alteration, trailed by the division of characters. In spite of a couple of positive tackles OCR over the world, improvement of OCR gadgets in Indian tongues is as yet a testing task. Character division expects a basic part in character affirmation since erroneously segmented characters are vulnerable to be seen wrongly. Hereafter the proposed work focuses on preprocessing and division of outdated interpreted reports. This is a basic walk towards making OCR for obsolete contents, which can be used by archeologists and curators for digitization and further examination of old records.

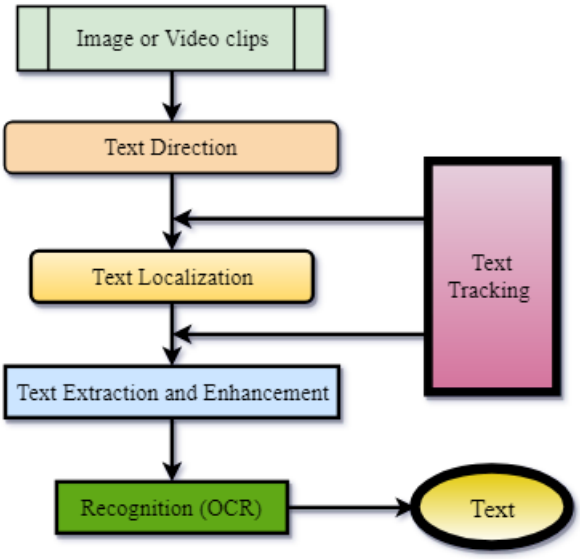

Figure 2 Text Extraction Process

It is done so as to improve the lucidity for human watchers of data in pictures, or to give better outcomes to the contribution of other mechanized picture handling strategies. Picture is improved by expelling clamor, fogginess. Picture smoothening is done to expel the impacts of camera clamor and missing pixel esteems. Picture honing is utilized to increase the detail that has been obscured because of commotion or different impacts, for example, unfocused camera, movement of article. Channels are utilized to expel clamor from the picture like middle channel, normal channel, low pass, and high pass channels (Fig. 2).

2. Literature Survey

[1]. Selvakumar, P., & Ganesh, S. H. (2017) presents an adaptable picture differentiate based record picture binarization technique that is tolerant to different sorts of record corruption, for instance, lopsided lighting up and file smear. The proposed strategy is direct and solid; simply a couple of parameters are incorporated. Likewise, it works for different sorts of defiled chronicle pictures. To completed watchful edge location calculation to look at and expel the words from a debased picture; the strategy passed on here is the image differentiate which is adaptively found to settle the issue. From the start, the separation directs is taken from the corrupted record pictures. The blend of close by picture point and the neighborhood picture contrast is the adaptable picture separation, and a while later it is changed over to twofold level and got together with watchful edge location calculation and fake neural system to focus content edge pixels. It makes usage of the neighborhood picture separate that is

evaluated considering the neighborhood most noteworthy and least and it has been taken a stab at the distinctive datasets. Tests show that the proposed system defeats most announced file binarization methods.

[2] Kavitha, B. R., & Srimathi, C. built up a CNN model without any preparation via preparing the model with the Tamil characters in disconnected mode and have accomplished great recognition results on both the preparation and testing datasets. An endeavor to set a benchmark for disconnected HTCR utilizing profound learning methods. To create a preparation precision of 95.16% which is much better contrasted with the customary methodologies. Displayed a profound learning approach for disconnected Handwritten Tamil Character Recognition. The outcomes have demonstrated that this methodology has gotten great execution when contrasted with the customary techniques. This testing exactness of 97.7% could even be improved by tuning the hyperparameters. Additionally, the greater part of the blunders was because of composing ambiguities of comparative characters. The outcomes displayed here can be utilized as a standard benchmark for HTCR.

[3] Akash V Pavaskar, Akshay S Accha, Anoop R Desai and Darshan K L (2017) proposed data extraction from pictures utilizing pytesseract and NLTK. To utilizing PC vision (Pytesseract) to remove helpful data like text, contact subtleties and hyperlinks from pictures. The android based application would enable the client to transfer a photograph and empower the client in putting away the contact subtleties, set a leftover portion, give a synopsis of the substance of the picture, the opening of hyperlinks legitimately from the application without expecting to type the URL inside the program. Accordingly, making the pictures an increasingly profitable and making the activity of the client all the more simple and advantageous. The extricating valuable substance from pictures is useful to the client. Utilizing Pytesseract to remove text from pictures and afterward characterizing it utilizing NLTK (Natural Language Toolkit). The data extraction from picture offers the chance to store certain subtleties including contact data, URLs, Date/Day in the arrangement client requires in order to be in a state of harmony with the quick-paced world. Along these lines, it coordinates different capacities under single application and diminishes intricacy of preparing and time.

[4] Manana Khachidze, Magda Tsintsadze, and Maia Archuadze (2016) the instrument for the arrangement of therapeutic records dependent on the Georgian language. It is the main endeavor of such characterization of the Georgian language-based therapeutic records. In general 24.855 assessment records have been considered. The archives were characterized into three primary gatherings (ultrasonography, endoscopy, and X-ray) and 13 subgroups utilizing two understood strategies: Support Vector Machine (SVM) and K-Nearest Neighbor (KNN). The outcomes got showed that both machine learning techniques performed effectively,

with a little matchless quality of SVM. During the time spent characterization a "contract" strategy, in light of highlights choice, was presented and applied. At the principal phase of order the consequences of the "recoil" case were better; be that as it may, on the second phase of grouping into subclasses 23% of all reports couldn't be connected to just a single unmistakable individual subclass (liver or paired framework) because of basic highlights characterizing these subclasses.

[5] Nitin Sharma and Nidhi give investigation and correlation of execution of different strategies utilized for the extraction of text data from pictures. Text extraction includes text discovery, restriction of text, following of text, extraction of text, upgrade, and text recognition from a given picture. Text discovery finds whether text is available in a given picture or not. Typically text discovery is applied for an arrangement of pictures. Text restriction limits the text inside the picture and bounding boxes are produced around the text. On the off chance that the text isn't situated in text confinement step, at that pointed text following is useful in finding that text. Text extraction alludes to separate text from pictures. Text extraction from pictures can be demonstrated valuable data for the content-based applications. So as to extricate the text, various strategies are utilized like area based-and texture-based strategy. Locale based strategies include associated part and edge-based techniques. Associated segment based strategy gives lackluster showing for consolidated characters or when the characters are not totally isolated from the picture foundation. The texture-based approach has a failure to perceive the characters that scope beneath the standard or above different characters, and these winds up in portioning a character into two parts. Edge-based technique likewise makes bogus expectation when the edge of any item in foundation of picture looks like any character.

3. Workflow

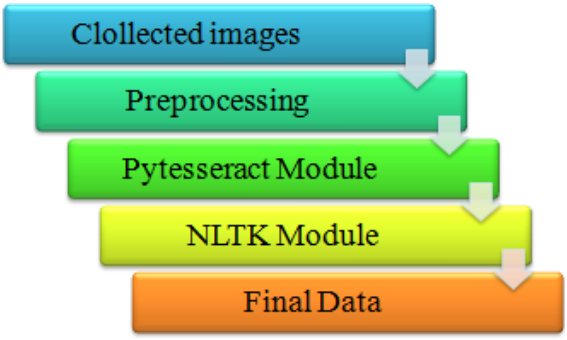

Figure 3 Workflow diagram

Stages

1. **Data Collection**

   Collection of pictures to be preprocessed, any picture having a place with single or different languages can be gathered and put away.

2. Preprocessing the images

All the gathered pictures are approved for required goals at that point changed over into .tiff group for better outcome.

3. Pytesseract

Each handled picture is given as a contribution to the pytesseract module and examined. The perceived text is printed/spared to a document thus.

Python-tesseract is a wrapper for Google's Tesseract-OCR Engine. It is additionally valuable as an independent conjuring content to tesseract, as it can peruse all picture types supported by the Pillow and Leptonica imaging libraries, including jpeg, png, gif, bmp, tiff, and others. Furthermore, whenever utilized as a content, Python-tesseract will print the perceived text as opposed to composing it to a document.

4. NLTK

Spared text subtleties are given as a contribution to the NLTK module and arranged dependent on language. Isolated language-based information is put away for handling.

Natural Language Toolkit

NLTK is the main stage for building Python projects to work with human language information. It gives simple to-utilize interfaces to more than 50 corpora and lexical assets, for example, WordNet, alongside a suite of text preparing libraries for ordering, tokenization, stemming, labeling, parsing, and semantic thinking, wrappers for modern quality NLP libraries.

Natural Language Processing with Python gives a handy prologue to programming for language handling. Composed by the makers of NLTK, it manages the per-user through the basics of composing Python programs, working with corpora, ordering text, breaking down etymological structure, and the sky is the limit from there.

5. Final Data

Subsequently, the assortment of pictures are changed over into process capable information which can be bolstered to information mining apparatuses and changed over into utilization full data.

Conclusion

The proposed technique is direct and solid; simply a couple of parameters are incorporated. What's more, it works for different kinds of defiled chronicle pictures. It makes use of the neighborhood picture separate that is surveyed considering the

neighborhood most prominent and least and it has been taken a stab at the diverse datasets. The accessible pictures in web gathered and prepared in the today techno world, by utilizing Big Data, AI and ML and so forth… In this diary, we simply proposing a technique which can in all likelihood be a one-stop answer for all.

References

1. Selvakumar, P., & Ganesh, S. H. "Tamil Character Recognition using Canny Edge Detection Algorithm" 978-1-5090-5573-9/16 $31.00 © 2016 IEEE.
2. Kavitha, B. R., & Srimathi, C. "Benchmarking on offline Handwritten Tamil Character Recognition using convolutional neural networks" 1319-1578. The Authors. Production and hosting by Elsevier.
3. Akash V Pavaskar, Akshay S Accha, Anoop R Desai and Darshan K L "INFORMATION EXTRACTION FROM IMAGES USING PYTESSERACT AND NLTK" May 2017, Volume 4, Issue 05. JETIR (ISSN-2349-5162).
4. Manana Khachidze, Magda Tsintsadze, and Maia Archuadze "Natural Language Processing Based Instrument for Classification of Free Text Medical Records" Volume 2016, Article ID 8313454, 10 pages. http://dx.doi.org/10.1155/2016/8313454.
5. Nitin Sharma and Nidhi "Text Extraction from Images: A Review" https://doi.org/10.1007/978-981-10-3920-1_16 © Springer Nature Singapore Pte Ltd.
6. Boufenar, C., Kerboua, A., Batouche, M., Investigation on deep learning for offline handwritten Arabic character recognition. Cognit. Syst. Res 2017.
7. Ciresan, D., Meier, U., Multi-column deep neural networks for offline handwritten Chinese character classification. 2015 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 1–6.
8. El-Sawy, A., Loey, M., Hazem, E.B., Arabic handwritten characters recognition using convolutional neural network. WSEAS Trans. Comput. Res. 5, 11–19. 2017.
9. K., Zhang, X., Ren, S., Sun, J., Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778. 2016.
10. Ioffe, S., Szegedy, C., Batch normalization: Accelerating deep network training by reducing internal covariate shift.arXiv preprint arXiv:1502.03167. 2015.
11. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C.,. Imagenet large scale visual recognition challenge. Int. J. Comput. Vision 115 (3), 211–252. 2015.
12. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., , June. Going deeper with convolutions. Cvpr. Tsai, C. Recognizing Handwritten Japanese Characters Using Deep Convolutional Neural Networks; Technical Report; Stanford University: Stanford, CA, USA, 2016; pp. 1–7. 2015.

13. Zhang, X.Y., Bengio, Y., Liu, C.L.,. Online and offline handwritten chinese character recognition: a comprehensive study and new benchmark. Pattern Recognit. 61, 348–360. 2017.
14. Viet Phuong Le, Nibal Nayef, Muriel Visani, Jean-Marc Ogier and Cao De Trant "Text and Non-text Segmentation based on Connected Component Features" IEEE, 13th International Conference on Document Analysis and Recognition (ICDAR), (2015).
15. Yingying Zhu, Cong Yao, Xiang Bai "Scene text detection and recognition: recent advances and future trends" Front. Comput. Sci., (2016).