WRITER RECOGNITION FOR SOUTH INDIAN LANGUAGES USING STATISTICAL FEATURE EXTRACTION AND DISTANCE CLASSIFIER

Aravinda C.V [1] and Dr. Prakash H.N[2]

[1]Department of Information Science & Engineering, SJBIT, Kengeri, Bengaluru

[2] Department of Computer Engineering, Rajeev Institute of Technology,Hassan

ABSTRACT

The comprehension of whole manually written records is a testing issue which incorporates various difficult undertakings. Given a written by hand archive, its format needs to be dissected to detach different content sorts in a first step. These different content sorts can then be coordinated to specific frameworks, including writer style, image, or table recognizers. Research in programmed author recognizable proof has principally centred around the measurable methodology. This has prompted the particular and extraction of factual elements, for example, run-length appropriations, incline dissemination, entropy, and edge-pivot conveyance. The edge-pivot conveyance highlight out flanks all other measurable elements. Edge-pivot circulation is an element that portrays the adjustments in bearing of a written work stroke in written by hand content. The edge- pivot circulation is extricated by method for a window that is slid over an edge-recognized on offline scanned images. At whatever point the focal pixel of the window is on, the two edge pieces (i.e. associated successions of pixels) rising up out of this focal pixel are considered. Their bearings are measured and put away as sets. A joint likelihood dissemination is gotten from an extensive recognition of such matches.

KEYWORDS

Euclidean distance , Similarity Edge detection, Text Detection

1. INTRODUCTION

Optical character acknowledgment (OCR) is one of the for all intents and purposes prominent uses of programmed example acknowledgment. Research in OCR is extremely well known for different application possibilities in banks, post workplaces, safeguard associations, perusing help for the visually impaired, library computerization, dialect preparing and multi-media plan. India is a multi-lingual multi-script nation, where a solitary archive page may contains content in two or more dialect scripts. OCR is of exceptional importance for a multilingual nation like India having 16 noteworthy state dialects and more than 100 provincial dialects. Character attestation can illuminate more identity boggling issues and energize the drudgery required in keeping up dull picture records. Fundamentally changing over enrolled image with substance report can draw in control through word get prepared applications. Optical Character Recognition has gotten an imperativeness since the essential for digitizing or changing over checked image of machine printed or physically made substance (numerals, letters, and image), in to a strategy saw by PCs, (for example, ASCII). OCR has been exhaustively utilized as the key use of various learning strategies in machine learning creating [1]. Penmanship assertion is the errand of changing a language re-appeared in its own particular spatial kind of graphical engravings into an average

DOI: 10.5121/ijnlc.2016.5403

representation [2]. written confirmation obtained diverse improvements from optical character insistence (OCR). The standard contrast amongst deciphered and typewritten characters is in the arrangements that run with writer style. It is in addition worth seeing that OCR regulates logged off insistence while writer style confirmation might be required for both on-line and withdrew from the net signs. Penmanship confirmation is one of the unequivocal testing issues. For the most part the field of writer style attestation is distributed into logged from time to time line insistence [2]. In isolated from the net insistence, as they say the photograph of the writer style is open for the PC, while in the on-line case transient data, for case, pentip empowers as a cutoff of time is in like way accessible. Commonplace information securing contraptions for logged every so often line insistence are scanners and digitizing tablets, independently. Because of the nonattendance of flitting data, withdrew from the net writer style attestation is seen as more troublesome than on-line. Additionally, it is moreover tidy that the logged up case is the one that takes a gander at to the standard investigating errand performed by people [3]. The essential for OCR creates concerning digitizing. The present difficulties in the documented of OCR innovation is to deal with low quality records, perceiving characters with different commotion sorts, acknowledgment of multilingual characters and advancement of OCR which can handle distinctive textual styles and sizes. For Indian and numerous other oriental dialects, OCR frameworks are not yet ready to effectively perceive printed archive image of changing scripts, quality, size, style, and textual style. Rather than European dialects, Indian dialects posture numerous extra difficulties. For example, (i) expansive number of vowels, consonants, and conjuncts, (ii) generally scripts spread more than a few zones, (iii) inflectional in nature and having complex character grapheme, (iv) absence of standard test databases for the Indian dialects.

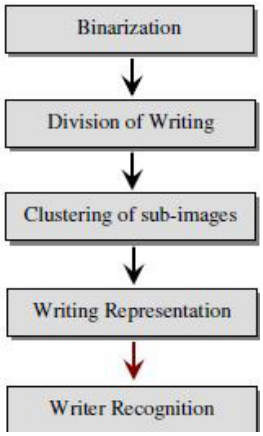

Figure 1: Block Diagram

2. OVERVIEW OF SOUTH INDIAN LANGUAGES

(i) **Kannada Script** :- Kannada is one of the real Dravidian dialects of southern India and one of the soonest dialects confirm epigraphically in India and talked by around 50 million individuals in the Indian conditions of Karnataka. The script has 49 characters in its Alpha syllabary and is phonemic. The Kannada character set is practically indistinguishable to that of other Indian dialects. The characters are characterized into three classifications: swaras (vowels), vyanjanas (consonants) and yogavaahas (part vowel, part consonants).

(ii) **Tamil Scirpt** : Tamil is a Dravidian dialect talked prevalently by Tamils in India and Sri Lanka, of speakers in numerous other nations. It is the official dialect of the Indian condition of Tamil Nadu, furthermore has official status in Sri Lanka and Singapore and having more than 7 million speakers. Tamil is one of the real dialects of the world. The Tamil script has 12 vowels, 18 consonants and five grantha letters. The script, be that as it may, is syllabic and not alphabetic. The complete script, in this way, comprises of the 31 letters in their autonomous structure, and an extra 216 combinant letters speaking to each conceivable mix of a vowel and a consonant

(iii) **Telugu Script**: Telugu, another Dravidian dialect talked by around 5 million individuals in the southern Indian condition of Andhra Pradesh and neighbouring states, furthermore in Bahrain, Fiji, Malaysia, Mauritius, Singapore and the UAE. Telugu is a syllabic dialect. Like most dialects of India, each image in Telugu script speaks to a complete syllable. Authoritatively, there are 18 vowels, 36 consonants, and three double images. Of these, 13 vowels, 35.

(iv) **Malayalam Script**: Malayalam is the dialect talked prevalently in the condition of Kerala, in southern India. It is one of the 23 official dialects of India, talked by around 37 million individuals. The dialect has a place with the group of Dravidian dialects. Both the dialect and its written work framework are firmly identified with Tamil; in any case, Malayalam has a script of its own. Malayalam dialect script comprises of 51 letters counting 16 vowels and 37 consonants. The prior style of composing is currently substituted with another style and this new script decreases the distinctive letters for typeset from 900 to under 90.

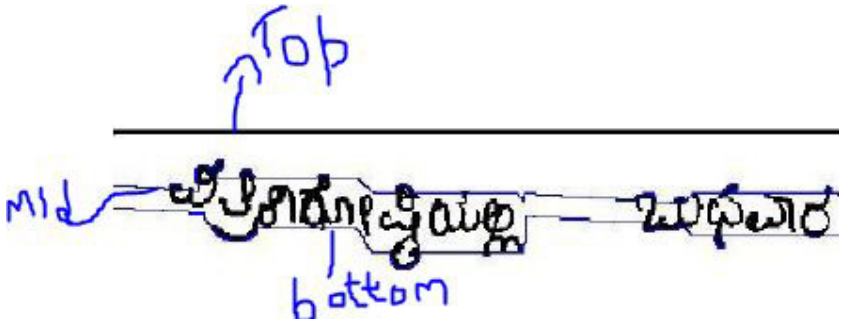

Figure 2: Segmentation of Kannada Handwritten Character.

Figure 3: Results Obtained after applying Edge Detection

3. PREVAILING WORKS SO FAR

This work exhibits an Offline Cursive Word Recognition System managing single essayist tests. The framework depends on a consistent thickness Hidden Markov Model prepared utilizing either the crude information, or information changed utilizing Principal Component Analysis or

Independent Component Analysis. Both procedures essentially enhanced the acknowledgment rate of the framework. Preprocessing, standardization and highlight extraction are portrayed and also the preparation method embraced. A few trials were performed utilizing a freely accessible database. The Malayam Letters precision acquired is the most elevated introduced in the writing over the same information. The framework depends on a sliding window approach: a window shifts section by segment over the picture and, at every progression, secludes an edge. A component vector is extricated from every edge and the grouping of edges so acquired is displayed with Continuous Density Hidden Markov Models (HMMs). The utilization of the sliding window approach has the vital point of preference of keeping away from the need of an autonomous division, a troublesome and blunder inclined procedure. Keeping in mind the end goal to decrease the quantity of parameters in the HMMs, we utilize corner to corner covariance grids in the emanation probabilities. This relates to the unreasonable presumption of having decorrelated highlight vectors. Consequently, we connected Principal Component Analysis (PCA) and Independent Component Analysis (ICA) to decorrelate the information. This permitted a critical change of the acknowledgment rate. The acknowledgment precision accomplished with the methodology proposed here is,to our insight, the most noteworthy among the outcomes over the same information displayed in the writing. The investigation of the acknowledgment as an element of the word length demonstrates that the framework accomplishes an acknowledgment rate for tests longer than six letters. This proposes the execution of our framework in errands including words with high normal length can be great. Both PCA and ICA positively affected the acknowledgment rate, PCA specifically diminished the mistake rate. A further change can most likely be acquired by utilizing nonlinear or bit PCA. Such strategies regularly work superior to the direct change we used to perform PCA. The utilization of information ward heuristics was maintained a strategic distance from keeping in mind the end goal to make the framework adaptable concerning a change of essayist. Any specially appointed calculation for the particular style of the author was maintained a strategic distance from. The earlier data about the word recurrence and appropriation can be helpful to enhance the acknowledgment of short words. These are normally articles, conjunctions and recommendations that show up frequently in the sentences. Consequently, a conceivable future course to take after is the use of dialect models that consider this sort of data.

Figure 4: Tamil Characters

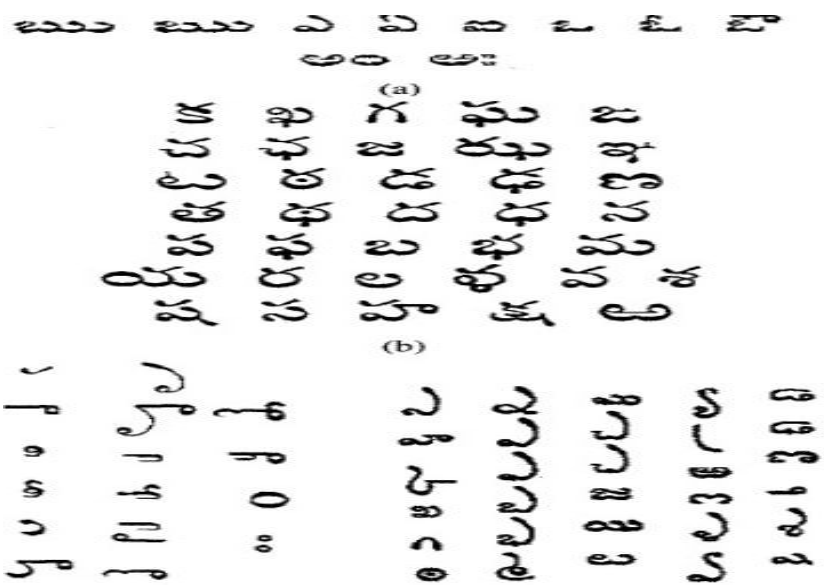

Figure 5: Telugu Characters

Figure 6: Malayalam Characters

IV. PROPOSED TECHNIQUES

In the proposed structure, this is proficient by a mix of names embeddings and properties learning, and a run of the subspace backslide. By then the images and strings talk to the same word which are close to each other allowing one to cast affirmation and recuperation assignments. The present system, advantage of our methodology has an changed to length, low dimensional and snappy to image. With difference to affirmation, composed by hand affirmation still speaks to a basic test for the same reasons. A model is at first arranged using stamped get ready data. At test time, given a word and a substance word, of that substance word being made by the model when maintained with the word. Affirmation can then be tended to by figuring the probabilities of all the vocabulary words and recuperating the nearest neighbor. As in the word spotting case, the major weakness here is the examination speed, since figuring these probabilities is solicitations of enormity slower than preparing an Euclidean partition or a spot thing between vectorial representations. The growing eagerness for removing abstract data. In any case, with

the late movement of gifted PC vision procedures some new frameworks have been proposed. To make feeling of how to recover and see words that have not been found amidst setting it up, is basic to have the capacity to exchange learning between the availability and testing tests.

Design and Acquisition : The reason for the database is in the first spot to serve as a dataset for the examination and advancement of techniques recognizing between different content sorts in online transcribed records. In the second place it thought to help with research about acknowledgment of record comment. The two necessities emerging in this manner are the need of different content sorts and the need that the records thought to be commented. To guarantee that every report contains sufficiently expansive measures of content, it has been chosen to control this by giving formats to the supporters which they needed to duplicate. This likewise tackled the issue of scholars not comprehending what to compose on the given white paper. An extra favourable position of utilizing layouts is the likelihood to control the substance conveyance.

5. PRE-PROCESSING

Preprocessing assumes a vital part in any OCR framework. In this segment we clarify two noteworthy preprocessing steps that are urgent for effective advancement of OCR framework:
(1) skew estimation and (2) division. The digitized pictures are in dark tone and we have utilized histogram based thresholding way to deal with proselyte them into two-tone pictures.

For an unmistakable report the histogram demonstrates two conspicuous crests relating to white and dark areas. The limit quality is picked as the midpoint of the two-histogram crests.The two-tone picture is changed over into 0-1 marks where the name 1 speaks to the article and 0 speaks to the foundation.

6. SEGMENTATION

By considering two issues identified with content comprehension: word spotting and word acknowledgment. In word recognizing, the objective is to discover all occasions of a question word in a dataset of image. The question word might be a content stringing which case it is normally alluded to as inquiry by string (QBS) or question by content (QBT) ,or may likewise be a picture, in which case it is generally alluded to as question by illustration (QBE). In word acknowledgment, the objective is to get a translation of the question word picture. By and large, including this work, it is accepted that a content word reference or vocabulary is supplied at test time, and that exclusive words from that dictionary can be utilized as competitor interpretations as a part of the acknowledgment undertaking. In this work we will likewise expect that the area of the words in the image is given, i.e., we have entry to image of trimmed words. On the off chance that those were not accessible, content confinement and division strategies could be utilized. In the division procedure we are editing the words indistinguishably and show it in the jumping box.

Recognition of Handwritten Words Original

The normalized version of this database consists of 250 samples from 50 writers. The samples written by 50 writers are originally used for training, cross validation and writer dependent testing, and the samples written by the other 14 are used for writer independent testing.

7. FEATURE SELECTION & EXTRACTION

The common stages of Feature Enhancement in Pattern Recognition field are Feature Selection & Feature Extraction. Feature selection techniques choose subsets of an originalset of features with two things: getting better classification rate by discarding irrelevant or poor features, or reducing the number of features as far as possible without decreasing the existing classification rate[5]. Feature extraction methods combine existing features in some way to create new features which better describe the input data.

1. Edge Direction Distribution

The first step, we identified the edge of the binary image using Sobel Detection method. These edge detected image are marked using 8 connected pixel neighbourhood. After this the number of rows & columns in an binary image is found using size function. Now the first black dot pixel in an image is identified & this pixel is considered as centre pixel of the square neighbourhood. Next we checked the black edge using the logical AND operator in all direction starting from the centre pixel & ending in any of the edge in square. To avoid the redundancy the upper two quadrants in the neighbourhood is checked & it is difficult to identify the direction of character written along the edge fragment. This will give the ”n” possible angles. These verified angles of each pixel are counted into n-binary histogram which is then normalized to a probability distribution which gives the probability of an edge fragment oriented in the image at the angle measured from the horizontal.

2. Feature Selection

In this we made the classification based on Collective Characters Feature Selection. The agenda we took to test, against zero , the difference μ_1 and μ _2 between the means of the valus taken by a feature in two classes. Let us take Class1 were xi; i = 1; 2; ::::;N; be the sample values of the feature in class w1 with μ_1: Likewise, for Class2 were w2 we shall take yi, i=1,2....,N with mean _2: For our assumtion the variance of the feature values is the same in both classes, μ_1 = μ_2 = μ_2 For the purpose of closeness of two mean values, hypothesis test was carried out.

$$H1 : \delta\mu = \mu1 - \mu2 \neq 0$$
$$H0 : \delta\mu = \mu1 - \mu2 = 0$$

Z=x-y

here x,y denote the random variables corresponding to the values of the feature in the two classes w1; w2 were statistical independence has been assumed. Likwise E[z]= μ1-μ2 and to the independence assumption $\sigma^2 z = 2\sigma^2$ We took the similar arguments were we used before, now we have this formula given below

$$\overline{z} = \frac{1}{N}\sum_{i=1}^{N}(xi - yi) = \overline{x} - \overline{y} \qquad (1)$$

and the known variance case follows the normal

$$N\left(\mu1 - \mu2, \frac{2\sigma^2}{N}\right) \qquad (2)$$

distribution for large N. In the table1 the values used to decide about the equation

$$\overline{z} = \frac{1}{N}\sum_{i=1}^{N}(xi - yi) = \overline{x} - \overline{y} \qquad (3)$$

If the variance is not known, then we choose the test static.

$$q = \frac{(\overline{x} - \overline{y}) - (\mu 1 - \mu 2)}{Sz\sqrt{\frac{2}{N}}} \qquad (4)$$

Where

$$S^2 z = \frac{1}{2N-2}\left\{\sum_{i=1}^{N}(xi - \overline{x})^2 + \sum_{i=1}^{N}(yi - \overline{y})^2\right\} \qquad (5)$$

(a) Height from baseline to upper edge:

By this system , we ascertained the tallness of the character from standard to upper edge is computed by deciding the gauge position of the character. This is finished by throwing an exhibit capacity here the file is line number in the character. Next the quantity of dark pixels in every line is calculated and the outcomes are put away in cluster. At long last the whole character ,the most extreme estimation of the cluster is perceived and the corresponding row number is put away as the baseline. The length of the character from the pattern to the upper edge is processed by subtracting the line number of first pixel in the picture from its column number of the gauge.

(b) Height from baseline to lower edge:

By this system the tallness of the twofold character from the pattern to the loweredge is controlled by figuring the benchmark column number. Next the column number of the last pixel of the character is considered. The stature of the picture from the benchmark to the lower edge is ascertained by subtracting the last pixel column number to the gauge line number.

(c) End Points:

End-points contain only one pixel in their 8-pixel neighbourhood. It is computed using end point function which gives the number of end points in the thinned image.

(d) Loop:

The loops of a character are the major distinguishing feature for many writers. The loop function gives loop length, angle of loop, position of the loop, area and average
radius of the loop of the edge image.

8. CLASSIFIERS

For this experiment we used the Manhanttan distance classifier & Eucledian Classifier. To calucate the variation of different angles this classifiers is very powerful.

For assumption equiprobable classes with the same covariance matrix, gx with the equation as shown below

$$gx = ln\left(p\left(\frac{x}{\omega i}\right)P\left(\omega i\right)\right) = lnp\left(\frac{x}{\omega i}\right) + lnP\left(\omega i\right) \qquad (6)$$

where constants have been neglected $\sum=\alpha^2 I$ In this case maximum gi (x) implies minimum. Euclidean distance .

de=$\| x-\mu i \|$ Thus featue vectors are asigned to classes according to their Euclidean distance from the respective mean points.

Distance Calculation

Distances were calculated for training and test set. For the extraction of the strings, a normalization of the curve was done. In one case, the string representation is given by the sequence of normalized segments, and in the other case by the sequence of angles between the segments.

Cost Function

The used cost functions are for the angle strings

$$c_k\left(\alpha \to \beta\right) = |\alpha - \beta|\left(angle - difference\right)$$
$$c_k\left(\epsilon \to \alpha\right) = k$$
$$c_k\left(\alpha \to \epsilon\right) = k$$

and for the segments (considered as vectors of a constant length)

$$c_k\left(\vec{x} \to \vec{y}\right) = |\vec{x} - \vec{y}|^k \; c_k\left(\epsilon \to \vec{x}\right) = c_k\left(\vec{x} \to \epsilon\right) = 2^{k-1}|\vec{x}|^k = 2^{k-1}l^k$$

9. Conclusion And Future Enhancement

This paper proposes an approach to manage address and consider word image for all South Indian Langauges like Kannada, Tamil, Telugu, Malayalam , both on record and on normal zones. We demonstrate how an attributes build approach based as for a pyramidal histogram of characters can be used to make sense of how to embed the word image and their printed elucidations into a common, more discriminative space, where the likeness between words is free of the composed work and content style, edification, get edge. This credits representation prompts a bound together representation of word image and strings, achieving a strategy that licenses one to perform either request by-delineation or inquiry by-string looks for, furthermore picture interpretation, in a united structure. We test our procedure in four open datasets of reports and trademark image, beating best in class approaches and showing that the proposed characteristic based representation is well suited for word looks, whether they are image or strings, in interpreted and ordinary image. The eventual outcomes of our philosophy on the word spotting undertaking. The edge-pivot conveyance highlight out flanks all other measurable elements. Edge-pivot circulation is an element that portrays the adjustments in bearing of a written work stroke in written by hand content. The edge-pivot circulation is extricated by method for a window that is slid over an edge-recognized on offline scanned images.

10. RESULTS OBTAINED

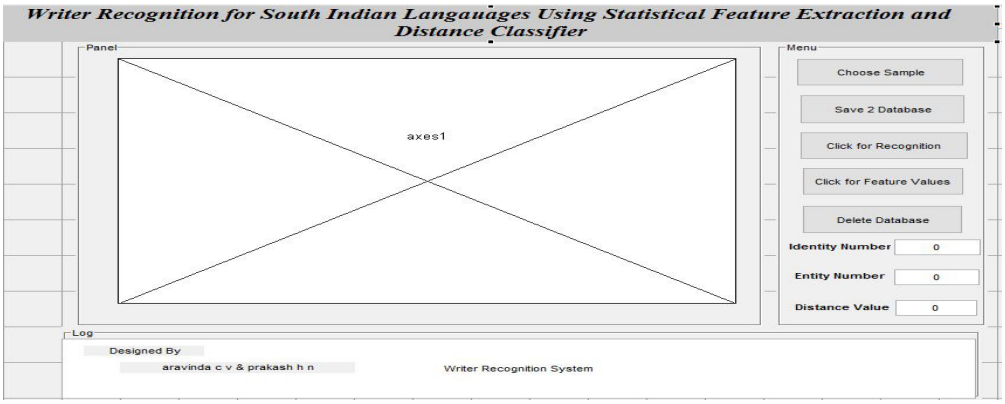

Figure 7 : Frame work for Testing

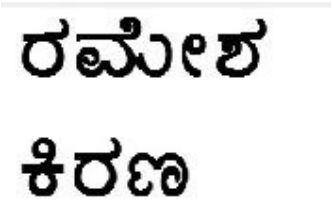

Figure 8 : Kannada Sample Input for Testing

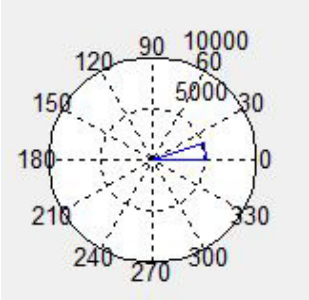

Figure 9: Polarized output for kannada

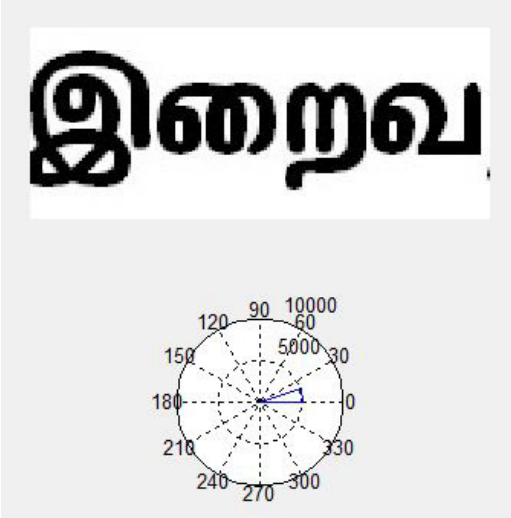

Figure 10 : Output for Tamil Sample

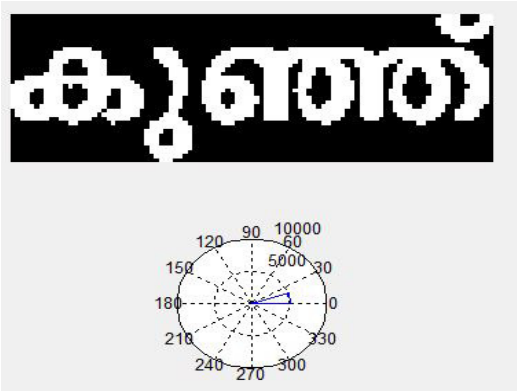

Figure 11 : Output for Malayalam Sample

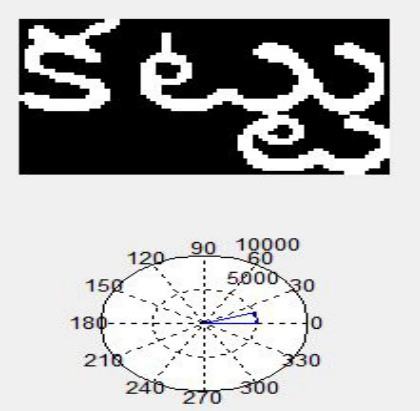

Figure 11: Output for Telugu Sample

ACKNOWLEDGEMENTS

The authors would like to thank everyone, just everyone!

REFERENCES

[1] Bhardwaj Anurag, Jose Damien and Govindaraju Venu. 2008. Script Independent Word Spotting in Multilingual Documents. In the Second International workshop on Cross Lingual Information Access-2008.

[2] B. Arazi. Handwriting identification by means of run-length measure- ments. IEEE Trans. Syst., Man and Cybernetics, SMC-7(12):878881, 1977

[3] L. R. B. Schomaker, A. J. W. M. Thomassen, and H.-L. Teulings. A computational model of cursive handwriting. In R. Plamondon, C. Y. Suen, and M. L. Simner, editors, Computer Recognition and Human Production of Handwriting, pages 153177. Singapore: World Scientific, 1989.

[4] L. Schomaker and L. Vuurpijl. Forensic writer identification: A bench- mark data set and a comparison of two systems [internal report for the Netherlands Forensic Institute]. Technical report, Nijmegen: NICI, 2000

[5] A. Raxiding, Research on Uyghur handwriting feature extraction method based on Gabor wavelet, (in Chinese), Journal of Hotan teachers college,(2010).

[6] L. Schomaker and L. Vuurpijl. Forensic writer identification: A bench-mark data set and a comparison of two systems [internal report for the Netherlands Forensic Institute]. Technical report, Nijmegen: NICI, 2000

[7] R. Plamondon and F. Maarse. An evaluation of motor models of handwriting. IEEE Trans. Syst. Man Cybern, 19:1060 1072, 1989

[8] L. R. B. Schomaker, A. J. W. M. Thomassen, and H.-L. Teulings. A computational model of cursive handwriting. In R. Plamondon, C. Y. Suen, and M. L. Simner, editors, Computer Recognition and Human Production of Handwriting, pages 153177. Singapore: World Scientific,1989

[9] J.-P. Crettez. A set of handwriting families: style recognition. In Proc. of the Third International Conference on Document Analysis and Recognition, pages 489494, Montreal, August 1995. IEEE Computer Society Press

[10] F. Maarse and A. Thomassen. Produced and perceived writing slant: differences between up and down strokes. Acta Psychologica, 54(1-3):131147, 1983

Authors

Aravinda cv currently working as Asst Prof at SJBIT college,kengeri, Bengaluru, Karnataka India. He has published 8 journals 7 Intenational papers and 3 National Level Conference papers.He has got 9 years of Teaching Experience at various Engineering Colleges. Currently he is pursuing his P.hD at VTU Belgavi.

Dr Prakash H.N Currently Professor and Head, Dept of Computer Science and Engineering Department at Rajeev Institute of Technology Hassan. He Completed his Ph.D(Computer Science) at Mysore University, M.Tech from N.I.T Warrangal, Hyderbad , B.E. at P.E.S(Mandya).He has got 24 years of Teaching in Engineering College. He has published 25 International Papers, 15 Journals and 15 National Level Conference papers. His area of interest Digitial Image Processing,Character Recognition,Signal Systems.