RECOGNITION OF MULTIPLE LANGAGES USING DIFFERENT FEATURE EXTRACTION METHODS

[1]Manjula S, [2]Poornima Nataraja

[1]Research Student, [2]Research Supervisor,
[1]Department of MCA,
[1]Dayananda Sagar College of Engineering, Bangalore,India.

Abstract : The main purpose of this research work is to recognize Tamil, Telugu and English languages present in multilingual document. As we know India is a multilingual country all official documents are having more than one language this type of document is known as multilingual document. Now world is moving towards paperless work all documents are going to be processed and saved in digital format, all the system is working towards auto identification of languages present in that document by considering some of the unique features of different languages. To achieve this task some of the statistical feature extraction techniques are implemented, such as; diagonal feature extraction, isomap feature extraction and parabola curve fitting feature extraction techniques. For the purpose of classification feed forward neural network classification model has been constructed this model is providing best recognition rate for all the three languages.

IndexTerms **- Multilingual document; diagonal feature extraction; isomap feature extraction; parabola feature extraction; feedforward neural network;**

I INTRODUCTION

Today India is working towards Digital India, our Honorable priminister declared e-governance system to create paperless work environment in future days where the entire text document will be maintained in electronic format. As we know India is a collection of multiple state each individual state is having their own standard regional language, hence India is also known as multi-lingual country. For the purpose of maintaining standardization in official communication Indian government has declared English as a international level and Hindi is a national level communication languages along with these two languages for state level communication standard regional language of that state is also declared as state level official communication language. These different regional languages belongs to different families of languages, among those language families major languages belongs to Indo-Aryan languages, that is 75% of Indians speaks Indo-Aryan languages this language has Devnagari script for writing purpose and 20% of Indians speaks Dravidian languages which has Dravidian script for writing purpose, where as remaining 5% of Indian uses Austro-Asiatic, Sino-Tibetan and isolates languages. According to 1961 census report of India there are 1652 different languages present in India. Whereas government of India has declared 22 languages as an official language. These official languages are Assamese, Bengali, Bodo, Dogri, Gujarati,Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam,Manipuri, Marathi, Nepali, Oriya,Punjabi, Sanskrit, Santhali, Sindhi, Tamil,Telugu and Urdu.The goal of Digital India project is to create paperless environment in all most all the fields, this is resulting towards automatic work functionality in all the government activities hence identification of languages present in each document is going to play a vital role in coming days.

This research work is focusing on Telugu, Tamil and English language identification and classification. Telugu is the regional language of Andrapradesh and Telangana state. Telugu language alphabets have 16 vowels, 3 vowel modifiers, 41 consonants and 60 symbols. Tamil is the regional language of Tamilnaadu state, Tamil language alphabets has 12 vowels, 18 consonants. These combine to form 216 compound characters and one special character and this combination creates total 247 characters [11]. Whereas English language alphabet has 5 vowels and 21 consonants, total 26 letters are there in English alphabet. For the purpose of automatic language identification some of the statistical feature extraction techniques such as; diagonal feature extraction, isomap feature extraction and parabola curve fitting feature extraction techniques are implemented, and for recognition neural network classification technique is implemented.

In the year 2011, Munish Kumar et.al., has implement diagonal feature extraction techniques on handwritten Gurumukhi script identification purpose[1]. In the same year J. Pradeep et.al, has implement diagonal feature extraction technique on handwritten English alphabets identification[10]. In the recent work Munish Kumar et.al, has implement parabola curve fitting feature extraction technique on handwritten Gurumukhi character identification[14]. The feature extraction techniques such as diagonal and parabola curve fitting feature extraction technique is only implemented on handwritten Gurumukhi and English character recognition purpose. In the recent work Qingbo Ji et.al, has implemented isomap feature extraction technique on signal processing signal recognition purpose [17]. This isomap feature extraction technique also used on palm vein verification purpose by Ali Mohsin Al-juboori et.al, in the year 2014[18]. In the year 2009, Xiao-li Xu has implemented isomap technique for sensitive feature extraction during fault prediction for electromechanical equipment [19]. Ming-Hsuan Yang has implemented extended isomap feature extraction technique, for face recognition and handwritten roman number recognition purpose [15].

In this research work diagonal, isomap and parabola curve fitting feature extraction technique are implement on printed document containing Tamil, Telugu and English languages identification, recognition is achieved by using neural network classification technique. The input image is having variable length of text document image, each individual input image is having different font

style, font size and there is no limitation of number of characters present in input image. This technique is flexible for different varieties of input image having Tamil, Telugu and English languages.

I. Literature Survey

Since many years numbers of researchers are working on identification of languages. Number of different techniques has been implemented to perform language identification task. In this paper we are presenting some of the techniques which were discussed by different researchers. In the year 2013, Rajneesh Rani et.al, discussed the zone based Gabor feature extraction technique for identification and recognition of Gurumukhi and English script with the help of SVM classifiers, this technique achieves average accuracy of 92.87% for linear kernel function, 93.28% for polynomial kernel function and 99.39% for Gaussian(RBF) kernel function[2]. In 2010 M.C Padma and P.A Vijaya has proposed texture-based approach to identify 7 south Indian languages, such as; Kannada, Tamil, Telugu, Malayalam, Urdu, Hindi and English, in this work the document images decomposed through wavelet packet using Haar basis function, K-NN classification technique has been implemented for the recognition of different languages, this method has achieved 99.68% of accuracy [3]. S.Hewavitharana and H.C.Fernando worked on recognition of handwritten Tamil characters using two stage classification, in first stage, unknown character is pre-classified into 3 different groups, that is core, ascending and descending groups, in 2nd stage the pre-classified characters are further analyzed by using horizontal projection profile feature extraction, based on this feature value pre-classified characters are identified and rate of recognition for this technique is 97% [4]. In the year 2011, Brijmohan Singh et.al, has worked for Devnagari handwritten character recognition by implementing curvelet transform and character geometry feature extraction technique, the extracted feature values are compared with SVM,RBF and KNN classification models, this has achieved 93.8% of accuracy[5]. Sk Md Obaidullah et.al, has worked on 6 different hand written languages such as Bangla, Devanagari, Malayalam, Urdu, Oriya and Roman identification, to perform identification task different feature extraction techniques were implemented such as mathematical features set, structural feature set and script dependent feature set, these technique has achieved 92.8% of recognition accuracy[6]. Priyank Mathur et.al, has worked on Bulgarian, Macedonian, Bosnian, Croatian, Serbian, Czech, Slovak, Peninsular Spain, Argentinean Spanish, Brazilian Portuguese, European Portuguese, Indonesian and Malaylanguages identification and these languages are recognized by Multinomial Naïve Bayes**,** Logistic Regression and Recurrent Neural Network, it has achieved 95.12% accuracy on discriminating between similar languages shared task[7]. Shijan Lu et.al, has worked on noisy and degraded document images language and script identification, to perform this task Lu has implemented different feature extraction techniques such as upward text boundary and lower text boundary technique, character extreme points identification, vertical direction movement of character or number identification with the help of position of the character or number[8]. In the recent year Sandhya Arora et.al, worked on recognition of non-compound handwritten Devnagari character using combination of MLP and minimum edit distance feature values, these values are used for classification two different MLP classifiers are used to get highest accuracy value, the overall recognition rated achieved for this technique is 90.74%[9]. In the year 2011, Abirami S et.al, has implemented tetra bit generation technique, in this method each character is segmented under 9 different zones and each zone will give unique value for each individual character, even head line analysis technique also implemented on Tamil, Hindi, English languages and English numerals, these extracted feature values are used for rule based classification technique for the purpose of classification of individual languages, this method has achieved 95.5% of accuracy for Tamil language recognition, 97% for English and 96% for Hindi and 94% for English number recognition[11]. In 2010, M.C.Padma and P.A.Vijaya has proposed a novel texture based approach to identify 10 different languages of printed document image, here image document is decomposed through wavelet packet decomposition technique, this feature identifies Bangla, Devnagari, English, Gujarati, Malayalam, Oriya and Tamil languages, K-NN classifier used for classification purpose and this technique achieved 98.24% of accuracy[13]. In the recent work N.Venkateswara Rao and B.Raveendra Babu has implemented chain code feature extraction technique on handwritten English digit identification and probabilistic neural network classification technique is used for the purpose of recognition, this method has achieved 98.1% of accuracy [16]. In the year 2011, Munish Kumar et.al., has implement zoning, diagonal, directional, intersection and open end points and Zernike moment feature extraction techniques on handwritten Gurumukhi script identification and K-NN, HMM and Bayesian classifier is used for classification purpose[1]. In the same year J. Pradeep et.al, has implement horizontal, vertical and diagonal feature extraction technique on handwritten document image containing English alphabets identification and neural network classifier technique has been implemented for recognition purpose, this technique has achieved 98.5% of accuracy [10]. In the recent work Munish Kumar et.al, has implement parabola curve fitting and power curve fitting feature extraction technique on handwritten Gurumukhi character identification, these feature values are used for recognition purpose by using SVM classifier. The maximum recognition rate for this technique is 89.12%[14]. The feature extraction techniques such as diagonal and parabola curve fitting feature extraction technique is only implemented on handwritten Gurumukhi and English character recognition purpose. In the recent work Qingbo Ji et.al, has implemented isomap feature extraction technique on signal processing to reduce high dimensional signal to low dimensional signal without lose of meaningful information present in the original signal, extracted low dimensional values are the features of the high dimensional signal which will be further used for signal recognition purpose[17]. This isomap feature extraction technique also used on palm vein verification purpose by Ali Mohsin Al-juboori et.al, in the year 2014. In his work he implemented isomap projection technique on Pam vein image samples; these input images are projected from high dimensional observation space to low dimensional through linear or nonlinear mapping to find out meaningful low dimensional values which are hidden in high dimensional data [18]. In the year 2009, Xiao-li Xu has implemented isomap technique for sensitive feature extraction during fault prediction for electromechanical equipment. In this paper he introduced isometric feature mapping algorithm based on the comprehensive analysis of the input data to reduce high dimensionality to low dimensionality data set, these extracted data set holds sensitive fault feature values of electromechanical equipment [19]. Ming-Hsuan Yang has implemented extended isomap feature extraction technique, in this technique the shortest path or distance between two points were identified, Yang implemented this extended feature extraction technique for face recognition and handwritten roman number recognition purpose[15].

As per rigorous literature survey, diagonal feature extraction technique, parabola curve fitting feature extraction technique and isomap feature extraction technique implementation done only for handwritten Gurumukhi scrip, English scrip and English number recognition purpose, here we can observe that both Gurumukhi and English languages are belongs to different scripts, this represents that each script is having unique features hence values after feature extraction is totally different for each individual scripts and it

will help for easy recognition of different scrip. Where as in this research work we are considering Telugu and Tamil languages both belongs to Dravidian script, challenging part of this paper is how efficiently these feature extraction technique will identify these languages with the help of neural network classification technique.

II. PROPOSED METHODOLOGY

The figure1 represents the proposed methodology for Telugu, Tamil and English language recognition.

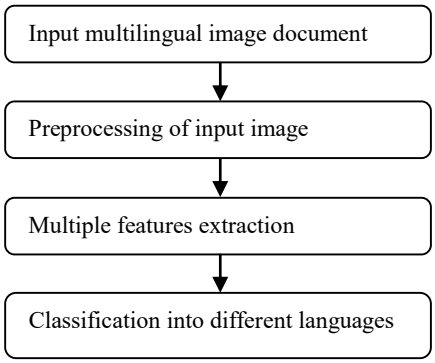

Figure1: Flow diagram of multilingual document identification and classification

Input Multilingual Image Document

The input image documents has been collected from different sources, most of the input images are color images and each image document is having minimum two languages, such as; Telugu – English language combination or Tamil-English language combination, hence these input text image documents are known as multilingual image document.

Preprocessing of Input Image Document

As we discussed in the previous stage most of the input images are color image documents, for the purpose of further operations these input color image documents need to convert into two dimensional image formats. This two dimensional image document is taken for the segmentation process. In this step each input image document is first undergo for line wise segmentation, each individual line is further undergo for word segmentation and each individual word is further undergo for character segmentation. During this preprocessing stage noise or irrelevant information present in the input image document are removed and valid data is going to be considered for further process. At next level skeletonization technique has been implemented on each individual segmented character. The resultant skeletonized image will be considered as an input image for feature extraction purpose.

Multiple Features Extraction

The major purpose of feature extraction technique is to minimization of memory utilization by extracting meaningful, valuable and unique information from set of input images. To perform this task different feature extraction techniques are implemented, such as; diagonal feature extraction, isomap feature extraction technique and parabola curve fitting feature extraction technique on each individual characters of Tamil, Telugu and English languages.

Diagonal Feature Extraction:The diagonal features are very important features in order to achieve higher recognition accuracy and reducing misclassification[1]. The features of the characters that are crucial for classifying them at recognition stage are extracted[10]. The prepocessed images has been standardized for the implementation of diagonal feature extraction. In this technique pixels are read in diagonal format and pixels with ON state will be added together for each individual diagonals present in an input image. The resultant sum value of each individual diagonal of an image are considered as a set of diagonal feature extracted values. In this experiment there are 47 different diagonal feature values are extracted for each individual character image.

Isomap Feature Extraction:Isomap is also known as Isometric feature extraction tecchnique. This technique uses multidimensional scaling and takes the distance between all points and calculates the position of each points. The main goal of isomap feature is to reduce high dimensional quality of data to low dimensional value by extracting meaningful values which are hidden in high dimensional data set. The problem of the dimensionality reduction process has received a more interest in many fields of information processing[18]. This techniques calculates the distance between pair of points, apartfrom this it also calculates pair wise distance between neighboring points. In this research work isomap feature extraction technique is implemented on standardized preprocessed image. The geodesic distance calulation method is implemented to calculate the distance between nearest neighbouring points of input image by taking threshold value. Landmark function is used to improve the speed and accuracy of this technique. This technique has generated 24 values from an input image, these 24 values are considred as low dimensional values. This extracted low dimensional information is also known as isomap feature extraction values, which will be further considered for recognition of languages.

Parabola Curve Fitting Feature Extraction:In this method standardized preprocessed image is segmented into 16 different blocks, each individual block will be used to caluclated parabola values. A prabola $y=ax^2+bx+c$ is uniquely defined by three paramenters a,b and c[14]. For each block fit a parabola, if any of the block dose not have any ON pixel then automatically assign a,b and c values as zero, otherwise implement lest square method to calculate value of a,b and c for each individual zone. This experiment has extracted 16 feature values for each individual character, The value extracted from this parabola curve fitting feature extraction is having unique set of paprabola values for each individual values. These set of values will reduce misclassification rate during the time of recognition of different languages.

Classification into Different Languages

The classification process will attempt to assign each input value to a given set of classes; this can be achieved by training a classification model with some set off input data set. In this research work neural network classification technique is used to

recognize Tamil, Telugu and English languages based on the values which are extracted from different feature extraction techniques. To obtain best recognition rate we are using supervised learning system by implementing feed forward neural network with sigmoid hidden neurons, here network can give consistent data and enough neurons in its hidden layer. During the construction of neural network we are concentrating on three different layers, such as; input layer, hidden layer and output layer. This constructed neural network is trained with Bayesian Regularization algorithm, this algorithm takes more time compare to all other algorithm, but it gives good recognition rate with very small error rate. This training system will stop according to adaptive weight minimization. The performance rate of this network is measured by calculating Mean Squared Error method. This Mean Squared Error is defined as the average squared difference between output and target values, during the time of calculation if we get zero value then it will be considered as no error or if we get lower values then it is considered as better performance. While training this neural network if we use variable number of hidden layers at each time then it will generate different resultant values based on number of hidden layers, which will affect even performance of the network. Even if network is training multiple times then also it will generate different resultant values due to different initialconditions and samplings. This trained neural network is used for the purpose of classification of test data.

III. EXPERIMENTAL RESULTS

The proposed system has been implemented on 1500 different noise free characters of Tamil, Telugu and English languages. These noise free character images are extracted from 200 different image documents. The characters present in these documents are having variable font type, size. The proposed operations are implemented by using MATLAB (R2015a). In this experiment three different feature extraction techniques has been implemented and each feature values are used for recognition of Tamil, Telugu and English languages by using neural network classification technique. To get maximum recognition rate the extracted feature values are sent as an input for neural network with 10 hidden layers, 25 hidden layers and 50 hidden layers. The recognition rates of each network for different feature extracted values are shown in below table.

	English	**Telugu**	**Tamil**	**Other**
English	412	09	69	10
Telugu	11	443	43	03
Tamil	52	35	413	00

Table 1: Diagonal Feature Recognition rate at Hidden Layer 10

	English	**Telugu**	**Tamil**	**Other**
English	470	05	15	10
Telugu	10	460	19	11
Tamil	13	06	473	08

Table 2: Diagonal Feature Recognition rate at Hidden Layer 25

	English	**Telugu**	**Tamil**	**Other**
English	351	25	11	13
Telugu	17	373	107	03
Tamil	42	99	359	00

Table 3: Diagonal Feature Recognition rate at Hidden Layer 50

Table 1,2, 3 shows the recognition of Tamil, Telugu and English languages at hidden layers 10, 25 and 50 by using diagonal feature extraction values. Some of the languages are misclassified for other groups. As per the observation hidden layer 25 is giving highest recognition rate such as English achieved 94.0% of accuracy, Telugu achieved 92% accuracy and Tamil 94.6% of accuracy with Mean Square Error of 0.020755 value.

	English	**Telugu**	**Tamil**	**Other**
English	314	21	162	03
Telugu	39	307	153	01
Tamil	61	46	393	00

Table 4: Isomap Feature Recognition rate at Hidden Layer 10

	English	**Telugu**	**Tamil**	**Other**
English	416	13	64	07
Telugu	26	431	39	04
Tamil	61	30	406	03

Table 5: Isomap Feature Recognition rate at Hidden Layer 25

	English	**Telugu**	**Tamil**	**Other**
English	445	09	29	17
Telugu	19	444	17	20
Tamil	61	08	463	13

Table 6: Isomap Feature Recognition rate at Hidden Layer 50

Table 4, 5, 6 shows the recognition of Tamil, Telugu and English languages at hidden layers 10, 25 and 50 by using isomap feature extraction values. Some of the languages are misclassified for other groups. As per the observation hidden layer 50 is giving highest recognition rate such as English achieved 89.0% of accuracy, Telugu achieved 88.8% accuracy and Tamil 92.6% of accuracy with Mean Square Error of 0.021153 value.

	English	**Telugu**	**Tamil**	**Other**
English	384	15	95	06
Telugu	22	371	105	02
Tamil	32	67	401	00

Table 7: Parabola Feature Recognition rate at Hidden Layer 10

	English	**Telugu**	**Tamil**	**Other**
English	410	12	63	15
Telugu	05	435	54	06
Tamil	20	40	439	01

Table8: Parabola Feature Recognition rate at Hidden Layer 25

	English	**Telugu**	**Tamil**	**Other**
English	448	08	26	18
Telugu	11	458	20	11
Tamil	07	20	465	08

Table9: Parabola Feature Recognition rate at Hidden Layer 25

Table 7, 8, 9 shows the recognition of Tamil, Telugu and English languages at hidden layers 10, 25 and 50 by using parabola feature extraction values. Some of the languages are misclassified for other groups. As per the observation hidden layer 50 is giving highest recognition rate such as English achieved 89.6% of accuracy, Telugu achieved 91.6% accuracy and Tamil 93% of accuracy with Mean Square Error of 0.029259 value.

IV. CONCLUSION

The main goal of this research work is to recognize Tamil, Telugu and English languages present in multilingual document using three different feature extraction techniques; such as, diagonal feature extraction, isomap feature extraction and parabola feature extraction technique. The neural network classifier model is used for the purpose of classification technique. As per the experiment, results shows that diagonal feature extraction technique is providing best recognition rate at hidden layer 25 with mean square error of 0.020755, isomap feature extraction technique is providing best result at hidden layer 50 with mean square error of 0.021153, whereas parabola feature extraction technique is providing highest recognition rate at hidden layer 50 with mean square error of 0.029259.

V. ACKNOWLEDGMENT

The first author extend her sincere gratitude to Department of MCA, Dayananda Sagar College of Engineering, VTU Research Center. for providing an opportunity to pursue my PhD work.

REFERENCES

[1] Manish Kumar, M.K. Jindal and R.K. Sharma, "Classification of characters and grading writers in offline handwritten gurumukhi script," Internaltional Conference on Image Information Processing(ICIIP) 2011. 978-1-61284-860-0/11.

[2]Rajneesh Rani, Renu Dhir and Gurpreet Singh Lehal, "Modified gabor feature extraction method for word level script identification- experimentation with gurumukhi and english scripts," International Journal of Signal Processing, Image Processing and Pattern Recognition. Vol. 6, No.5. 2013.

[3] M.C. Padma and P.A. Vijaya, "Global approach for script identification using wavelet packet based features," International journal of Signal Processing, Image Processing and Pattern Recognition. Septer 2010. Vol.3, No.3.

[4]S.Hewavitharana and H.C. Fernando, "A two stage classifiction approach to tamil handwritting recognition," University of Colombo. Colombo-03.

[5] Brijmohan Singh, Ankush Mittal and Debashis Gosh, "An evaluation of different feature extractors and classifiers for offline handwritten devanagari characters recognition," Journal of Pattern Recognition Research2 -2011. Pp:269-277.

[6]Sk Md Obaidullah, Supratik Kundu Das and Kaushik Roy, "A system for handwritten script identification from indian document," Journal of Pattern Recognition Research8-2013.

[7]Priyank Mathur, Arkajyoti Misra and Emrah Budur, "Language identification from text documents", Stanford university, California.

[8] Shijan Lu and Chew Lim Tan, "Script and language identifiation in noisy and degraded document images," Pattern Analysis and Machine Intelligence-2008.

[9] Sandhya Arora, Debotosh Bhattacharjee, Mita Nasipuri, D.K.Basu and M. Kundu, "Recognition of non-compound handwritten devanagari characters using combination of mlp and minimum edit distance," International Journal of Computer Science and Security(IJCSS). Vol.4, Issue-1.

[10] J.Pradeep, E.Srinivasan and S.Himavathi, "Diagonal based feature extraction for handwritten alphabets recognition system using neural network," International Journal of Computer Science and Information Technology(IJCSIT)-2011. Vol-3. No-1.

[11] Abirami S and Murugappan S, "Scripts and numerals identification from printed document images," David Bracewell et.al, AIAA-2011, CS&IT-03, pp:129-146, 2011.

[12] M.C. Padma and P.A. Vijaya, "Wavelet packet based textire feature for automatic script identification," International Journal of Image Processing-2010. Vol-4, Issue-01, pp 53-65.

[13] Hiremath P.S, Shivashankar S, Jagdeesh D. Pujari and V.Mouneswara, "Script identification in handwritten document image using texture feature," IEEE 2nd International Advance Computing Conference-2010. 978-1-4244-4791-6/10.

[14] Munish Kumar, R.K. Sharma and M.K. Jindal, "Size of training set vis-à-vis recognition accuracy of handwritten character recognition system," Journal of Emerging Technologies in Web Intelligence. Vol.5, No.4.2013, pp:380-384.

[15] Ming-Husan Yang, "Extended isomap for pattern classification," American Association for Artificial Intelligence.

[16] N.Venkateswararao and B.Raveendrababu, "Combined histogram chain code feature extraction method to recognize handwritten digits with probabilistic neural network," International Journal of Applied Engineering Research. ISSN 0973-4562, Vol-9, No.20. 2014.

[17] Qingbo Ji, Boyang Feng, Yun Lin, Zheng Dou, Zihiqiand Wu and Zhiping Zhang, : Indentification of digital signals based on manifold learning," International Journal of Signal Processing, Image Processing and Pattern Recognition, Vol-9, No-2, 2016. pp:127-134.

[18] Ali Mohsin Al-juboori, Wei Bu Xiangqian Wu and Qiushi Zhao, "Pam vein verification using multiple features and isometric projection," International Journal of Signal Processing, Image Processing and Pattern Recognition. Vol-7, No.1 2014.

[19] Xiao-li Xu, "ISOMAP algorithm based feature extraction for electromechanical equipment fault prediction," Image-Signal Processing-2009. IEEE publisher.2009. ISBN 978-1-4244-4129-7.