ABSTRACT

Writer identification is the process of identifying the individual based on their handwriting. Handwriting exhibits behavioral characteristics of an individual and has been considered as unique. The style and shape of the letters written vary slightly for same writer and entirely different for different writers. Hence accurate prediction of individual based on his/her handwriting is highly complex and challenging task. Writer identification is of a primordial importance in forensic expert decision-making systems, signature verification system, biometric authentication in information and network security.

Writer identification is an important research area undergone for past three decades. Writer identification has been studied extensively for several languages like Chinese, English, Arabic, Urdu and also for several Indian languages like Hindi, Oriya, Bangla, Kannada and Malayalam. Various approaches like pattern matching, statistical, machine learning have been adopted for processing the documents and developing the writer identification models for these languages.

The primary focus of this thesis titled "Performance Enhancement of Classifiers for Tamil Writer Identification through Modified Support Vector Machine (SVM) Linear Kernel with Parameter Estimation and Deep Learning" is to provide an efficient machine learning based solution for predicting accurately the identity of an individual based on his/her handwriting written in Tamil language. The writer regonization task is modelled as pattern classification and implemented based on offline text dependent mode using shallow and deep learning approaches. The major objectives of this research are,

- To develop a framework based on Supervised Learning Approach for Tamil writer identification
- To identify and capture discriminative features from Tamil handwriting and to build SVM classifiers
- To enhance the performance of SVM linear kernel by defining new linear kernels using parameter estimation techniques like Weighted Least Square Regression, Bayesian Linear Regression and Principal Component Regression
- To build SVM classifiers with new linear kernels such as Weighted Linear Kernel, Bayesian Linear Kernel and Principal Component Kernel
- To build Convolutional Neural Network - a Deep Learning framework for Tamil Writer Identification

- To develop a writer identification tool for identifying an individual / a person based on their Tamil handwriting.

The thesis explains the modeling of writer identification problem as classification task. Handwriting image classification is done based on both hand crafted features and self-extracted features through supervised learning and deep learning respectively. These approaches for writer identification exceedingly simplify the traditional writer identification problem and the prediction model is more effective, reliable since it is generated based on intelligent hints collected from the handwriting of individuals.

In modeling automatic writer identification, the essential tasks such as corpus preparation, preprocessing, feature extraction, training and writer prediction have been carried out to produce more accurate identification results.

The work is carried out for three levels of text (i) character (ii) word (iii) paragraph. As there is no benchmark datasets available online for Tamil handwriting text images, real time datasets have been created. To create datasets, the writers of different age groups are considered and the individuals have been identified based on various factors like education level, age, gender and locality.

A handwritten document with 100 paragraphs in 20 pages is designed and prepared by considering the intricacies involved in Tamil alphabets/characters. These text dependent documents written by 300 individuals are collected and scanned to convert into digital images. The paragraph text image is segmented into words and then characters. 100 words are chosen for each writer for word level processing and 100 characters are chosen for each writer for character level processing. Finally three corpuses have been developed.

The images in three different corpuses are preprocessed separately using various image processing techniques. The preprocessing tasks carried out on the character and word images are binarization, thresholding, dilation, edge detection and thinning. In case of paragraph text, the images are normalized and converted into grayscale images to carry further preprocessing tasks. The preprocessing tasks such as edge detection, image dilation and box bounding are carried out to extract highly discriminative features.

Feature extraction is another important task in modeling writer identification. In order to recognize the individual based on his/her handwriting, it is essential to define and capture the descriptive features from the handwritten text. In character and word text images, global features or textural features are derived using gray level co-occurance matrices and various structural properties

like word measurement features, morphological features, and fractal features of the handwriting are derived as local features. In case of paragraph text image, features such as Gabor Filter, Gray Level Co-Occurrence Matrix (GLCM), Generalized Gaussian Density (GGD), Contourlet GGD, and directional features were computed from pre-processed images.

Three independent datasets named TWINC, TWINW and TWINP have been developed using features related to character, word and paragraph text images respectively. In case of character and word text, the size of feature vector is 26 and the feature vectors are assigned class labels from 1 to 300 (as the number of writers is 300) which are considered as identity of persons. In paragraph text a total of 2784 features have been derived and a dimensionality reduction technique namely particle swam optimization feature selection method is used to select the well contributing features, which finally forms a feature vector of size 422. In all three cases, the datasets containing 30000 feature vectors are normalized using min-max normalization in order to make the feature values to lie within a specified range (0-3).

In the first experiment, Support Vector Machine (SVM) based models are developed using linear, polynomial and RBF kernels for multi class classification. The efficient classifiers are built using three datasets by tuning the regularization, degree, gamma parameters. The predictive performance of the classifiers is evaluated using various metrics like predictive accuracy, precision, recall, F-measure, time taken.

The SVM implementation with various kernels showed that linear kernel attains very low accuracy compared to other two kernels. But the observations showed that linear kernel performs faster than the other kernels with less computational complexity. Hence, a modified linear kernel is proposed to enrich the performance of the linear kernel using parameter estimation technique. In the second work Weighted Least Square (WLS) regression is used to estimate the dot products of the linear kernel and the kernel is referred as WLK. SVM with new form of linear kernel WLK is implemented with three different datasets and the results are analyzed.

The work is extended to develop another linear kernel using Bayesian Linear Regression and promoted the significance of linear kernel. BLR is a parameter estimation method based on Bayesian inference statistical analysis that is used to estimate the regression parameters. These parameters will act as the co-efficients of the Bayesian Linear Kernel (BLK). SVM with BLK kernel has been implemented for three datasets by tuning C- regularization parameter and the performances are evaluated.

In the next work another linear kernel is proposed using Principal Component Regression (PCR) which determines principal components that are used to estimate the coefficients of the linear kernel and this kernel is referred as PCK. SVM with PCK linear kernel is implemented for three datasets and the classifiers are evaluated. The proposed PCK-SVM shows comparatively higher performance when compared to WLK-SVM, BLK-SVM classifiers.

Convolutional Neural Networks (CNN) in Deep Learning have its own dimension to generate new features from a limited set of training dataset. The performance of CNN in extraction of high-level features is very efficient in dealing with shape changes that will probably be the key challenge in coping up with too much of cursiveness in Tamil writings. Hence in the next work CNN is employed to develop the writer identification models using the handwritten text corpuses of three categories character, word and paragraph. The performance of the CNN based classifiers is evaluated in terms of accuracy, precision, recall and F-measure. Justifiably, the CNNs produced much higher identification rate compared to traditional Artificial Neural Network.

Finally, an interactive discriminative writer identification tool has been developed using MATLAB for predicting the identity of writer based on their Tamil handwriting. The writer identification model based on CNN is incorporated for developing the tool. Two independent classifiers (i) word text image trained model (ii) paragraph text image trained model, are integrated into a single writer prediction system, which can accept a Tamil handwritten text either a single character or word or a paragraph text as input and can predict the writer of the handwriting.

As it is imperative to develop accurate models to identify the individual based on behavioral characteristics, in few important applications, this research work has been launched. Exhaustive experimentations carried out on Tamil handwritings ascertain that the classification modeling is effective in predicting the identity of writers.