Introduction

1. INTRODUCTION

1.1. OVERVIEW

Image classification is an area in image processing where the primary goal is to separate a set of images according to their visual content into one of a number of predefined categories. It is the problem of finding a mapping from images to a set of classes, not necessarily object categories. Each class is represented by a set of features (feature vector) and the algorithm that maps these feature vectors to a class uses machine learning techniques. The ability to perform multi-class image classification as an automatic task using computer is increasingly becoming important. This is due to the huge volume of image data available, which are proving to be difficult for manual analysis. The difficulty arises because of lack of human experts, poor quality images and time complexity. The current market need is to have techniques which can classify images with minimum intervention from the users in an efficient and effective manner.

This research work is an attempt made to develop such a classification system using machine learning algorithms. In particular, the study attempts to classify Tamil Isolated Handwritten Character (TIHC) images into groups. The strength of the selected feature and the effectiveness of the classifier are the two key factors determining the performance of a handwritten Character Recognition System. In the present research work, a fusion classification method that uses the visual features of the TIHC image and machine learning classifiers are proposed.

This chapter provides an introduction to the concepts of various topics that related to this research work. Section 1.2 provides an introduction to Indian scripts with emphasize to Tamil characters. Section 1.3 describes the components of OCR system. A brief introduction to the image features and classifiers is given in Section 1.4; Section 1.5 presents the various factors that motivated the present

research work, while Section 1.6 presents the problem statement along with the formulation research objectives.

1.2. OPTICAL CHARACTER RECOGNITION (OCR)

The main application of pattern recognition concepts is made in the field of 'Character Recognition', which has major implications in automation and information handling. Optical character recognition (OCR) systems consist of a scanning device and pattern recognition software that translates the scanned imaged into computer coded characters.

It is the process by which a computer analyzes a static image of a document and translates the words within the image into text characters. The text can then be modified, searched, or copied as in a standard text document. OCR technology is now employed in a wide variety of fields to digitize documents normally received or maintained in hard copy.

The various components of OCR system are.

(i) Data Acquisition
(ii) Preprocessing
(iii) Segmentation
(iv) Feature Extraction
(v) Classification

Data acquisition normally consists of a digitized image of the document containing the characters to be recognized and obtained using an optical scanner. After the regions containing the characters are located, each character is extracted through a segmentation process. The extracted digits are followed, in most of the cases, by a preprocessing, so as to eliminate noise and to facilitate the extraction of features in the next step. Feature extraction is one of the most important step in the

recognition system, since the feature selected has to represent well the pattern which as to be recognized. In many case, the amount of data selected in the feature extraction is huge so a reduction of this data is necessary.

The last stage of OCR is classification, which uses the extracted features for classifying the characters. The classification of OCR system is composed of two phases, namely, training and testing phases. During the training phase, a set of characters from the character image database are preprocessed, segmented and their features are extracted. These features are used to train the classifier. Later, during testing, when a new character pattern is acquired, the features extracted through the same process are compared against the training set and a match process is performed. The character which has the highest matching score is declared as the recognized digit.

The present study focuses on improving the last step of OCR, that is, classification process. For this purpose, a fusion based classification algorithm is proposed using a combination of Neural Network, K-Nearest Neighbor and Support Vector Machines. The OCR is performed for handwritten Tamil characters. In order to better understand the features of the image dataset, a brief description of the Tamil Script is provided in the following section.

1.3. TAMIL SCRIPT

Indian scripts are generally written in non-cursive style, unlike Latin alphabet, which is normally written in cursive style, rendering recognition difficult. However, Indian scripts pose a peculiar problem nonexistent in European scripts, that is, the problem of composite characters. Unlike Latin alphabets where a single character represents a consonant or a vowel, in Indian scripts a composite character represents either a complete syllable, or the coda of one syllable and the onset of another. Therefore, although the basic units that form composite

characters of a script are not that many ($O(10^2)$), these units by various combinations lead to a large number ($O(10^4)$) of composite characters.

Tamil language belongs to the Dravidian scripts of Southern India. It is one of the classical languages of the world with a literary history of more than two millenniums spanning from the Cankam age (300 BC – 200 AD). Tamil is one of the oldest language has several million speakers across the world and is an official language in countries such as Sri Lanka, Malaysia, Singapore and Tamil Nadu State of India. As it is the case with all Indic scripts, Tamil has a large alphabet size and hence text entry through QWERTY keyboard is cumbersome. The penetration of Information Technology (IT) becomes harder in a country such as India where the majorities read and write in their native language. Therefore, enabling interaction with computers in the native language and in a natural way such as handwriting is absolutely necessary.

The Tamil script has twelve vowels, eighteen consonants and one character, the aytam, which is classified in Tamil grammar as being neither a consonant nor a vowel, though often considered as part of the vowel set. The script is syllabic and not alphabetic (Iravatham, 2003). The complete script consists of the thirty-one letters in their independent form (Figure 1.1), and an additional 216 combinatory letters (Figure 1.2) representing a total 247 combinations of a consonant and a vowel, a mute consonant, or a vowel alone. These combinant letters are formed by adding a vowel marker to the consonant. Some vowels require the basic shape of the consonant to be altered in a way that is specific to that vowel. Others are written by adding a vowel-specific suffix to the consonant, yet others a prefix, and finally some vowels require adding both a prefix and a suffix to the consonant. In every case the vowel marker is different from the standalone character for the vowel.

Type	No. of Scripts	Scripts
Vowels and Ayutha Letter	12	அ ஆ இ ஈ உ ஊ எ ஏ ஐ ஒ ஓ ஔ ஃ
Consonants	18	க ங ச ஞ ட ண த ந ப ம ய ர ல வ ழ ள ற ன
Consonants from Sanskrit	6	ஜ ஶ ஷ ஸ ஹ க்ஷ

Figure 1.1: Tamil Scripts

Vowels → Consonants ↓	அ	ஆ	இ	ஈ	உ	ஊ	எ	ஏ	ஐ	ஒ	ஓ	ஔ
க்	க	கா	கி	கீ	கு	கூ	கெ	கே	கை	கொ	கோ	கௌ
ங்	ங	ஙா	ஙி	ஙீ	ஙு	ஙூ	ஙெ	ஙே	ஙை	ஙொ	ஙோ	ஙௌ
ச்	ச	சா	சி	சீ	சு	சூ	செ	சே	சை	சொ	சோ	சௌ
ஞ்	ஞ	ஞா	ஞி	ஞீ	ஞு	ஞூ	ஞெ	ஞே	ஞை	ஞொ	ஞோ	ஞௌ
ட்	ட	டா	டி	டீ	டு	டூ	டெ	டே	டை	டொ	டோ	டௌ
ண்	ண	ணா	ணி	ணீ	ணு	ணூ	ணெ	ணே	ணை	ணொ	ணோ	ணௌ
த்	த	தா	தி	தீ	து	தூ	தெ	தே	தை	தொ	தோ	தௌ
ந்	ந	நா	நி	நீ	நு	நூ	நெ	நே	நை	நொ	நோ	நௌ
ப்	ப	பா	பி	பீ	பு	பூ	பெ	பே	பை	பொ	போ	பௌ
ம்	ம	மா	மி	மீ	மு	மூ	மெ	மே	மை	மொ	மோ	மௌ
ய்	ய	யா	யி	யீ	யு	யூ	யெ	யே	யை	யொ	யோ	யௌ
ர்	ர	ரா	ரி	ரீ	ரு	ரூ	ரெ	ரே	ரை	ரொ	ரோ	ரௌ
ல்	ல	லா	லி	லீ	லு	லூ	லெ	லே	லை	லொ	லோ	லௌ
வ்	வ	வா	வி	வீ	வு	வூ	வெ	வே	வை	வொ	வோ	வௌ
ழ்	ழ	ழா	ழி	ழீ	ழு	ழூ	ழெ	ழே	ழை	ழொ	ழோ	ழௌ
ள்	ள	ளா	ளி	ளீ	ளு	ளூ	ளெ	ளே	ளை	ளொ	ளோ	ளௌ
ற்	ற	றா	றி	றீ	று	றூ	றெ	றே	றை	றொ	றோ	றௌ
ன்	ன	னா	னி	னீ	னு	னூ	னெ	னே	னை	னொ	னோ	னௌ

Figure 1.2: Combinant Tamil Letters

While consideration the task of classifying Tamil characters, each character is considered as a class and thus becomes a multi-class problem. This research work focus on designing such a classifier.

1.4. MACHINE LEARNING

Machine learning is the process of automating the development of some part of a system which performs some task. The algorithm, parameters to an algorithm or process can be learnt adaptively over a period of time. The overall structure of a machine learning approach to a problem involves three steps (Russell and Norvig, 2003).

1. The generation of some representation of a solution to the problem.
2. The evaluation of the generated solution.
3. If the evaluated solution is not good enough, the solution is iterated (i.e. almost always improved) and the machine learning process goes to step 2.

1.4.1. Types of Generation and Representation

Given the selection of some representation of a solution to the problem, the initial generation is usually random but constrained by some parameters. For example, in a Neural Network the structure is fixed and the weight associated with each link is generated according to some algorithm which ensures that the initially generated solution will almost certainly not be the same from time-to-time. However, there are a wide variety of possible representations, including

(i) Feed-Forward Neural Networks

(ii) Genetic Algorithms

(iii) Support Vector Machines

(iv) Simulated Annealing

(v) Decision Trees

(vi) Bayesian Networks and

(vii) Genetic Programming.

1.4.2. Types of Iteration

There are, broadly, three ways in which the iteration from one solution to the next can be performed. This iteration is how the search for a good solution is carried out. Each of the three types of iteration will be summarized briefly in this section.

(i) Unsupervised - Unsupervised learning is normally used to locate patterns in the input data. No information is given to the system, which finds the patterns as to the correctness or incorrectness of the patterns. The patterns it finds may therefore be arbitrary or they may actually be representative of some real underlying process which caused them to appear gives for details on unsupervised classification problem.

(ii) Reinforcement - Reinforcement in terms of the quantity of information given to the system regarding the correctness of its output. Reinforcement learning is intermediary between supervised and unsupervised learning. When reinforcement learning is used some information is given to the system at some time regarding the correctness of a prediction it made. This information ranges in precision from a categorization of a response as "right" or "wrong" to a precise amount of error, expressed numerically. At the later degree of precision it differs from supervised learning only in the way the information is presented.

(iii) Supervised - When supervised learning is used the precise, correct output which should have been given for any particular training input is known to the system and used by the system to adjust the answer it will give to other training examples.

1.4.3. Types of Evaluation

The performance of the classifiers can be determined using various performance parameters like accuracy, speed and error rate. The types and ways in which evaluation can be carried out are discussed in Chapter 3.

In the present research work various machine learning classifiers are combined to classify the TIHC images.

1.5. GENERAL APPROACH TO CLASSIFICATION

As mentioned earlier, classification, also known as pattern recognition, discrimination, supervised learning or prediction, is a task that involves construction of a procedure that maps data into one of several predefined classes (Montejo-Raez, 2005). It applies a rule, a boundary or a function to the sample's attributes, in order to identify the classes. Classification can be applied to databases, text documents, web documents, web based text documents, etc. Classification is considered as a challenging field and contains more scope for research. Some example applications include predicting tumor cells as benign or malignant, classifying credit card transactions as legitimate or fraudulent, classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil, categorizing news stories as finance, weather, entertainment, sports, etc. and grouping e-mails as spam or non-spams. It is considered challenging (Wu, 2006) because of the following reasons

- Information overload – The information explosion era is overloaded with information and finding the required information is prohibitively expensive.
- Size and Dimension – The information stored is very high, which in turn, increases the size of the database to be analyzed. Moreover, the databases have very high number of "dimensions" or "features", which again pose challenges during classification.

A basic classification model is shown in Figure 1.3.

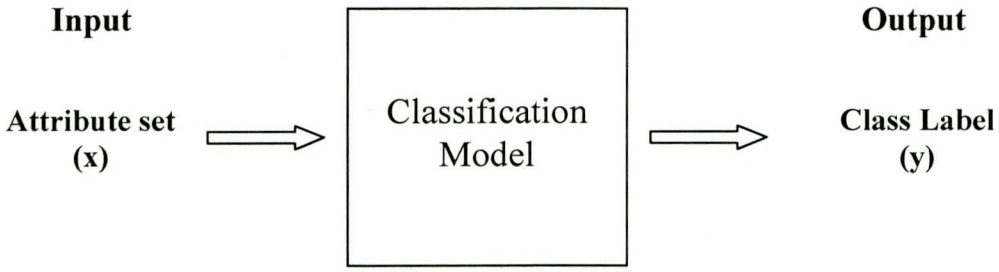

Figure 1.3: Classification Model

The input data for a classification task is a collection of features arranged as in row-wise fashion (records). Each record, also known as an instance or example, is characterized by a tuple (X, y) where X is the attribute set and y is a special attribute, designated as the class label (also known as category or target attribute).

1.5.1. Definition

Classification is the task of learning a target function f that maps each attribute set X to one of the predefined class labels y. The target function is also known informally as a classification model and is useful for the following purposes.

1.5.2. Classification and Prediction

Classification is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts, for the purpose of being able to

use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known).

The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae or Neural Networks. Out of these, the use of decision trees representation is more popular as they can be easily converted to classification rules.

Classification can be used for predicting the class label of data objects. However, in many applications, users may wish to predict some missing or unavailable data values rather than class labels. This is usually the case when the predicted values are numerical data and is often specifically referred to as prediction. Although prediction may refer to both data value prediction and class label prediction, it is usually confine to data value prediction and thus is distinct from classification. Prediction also encompasses the identification of distribution trends based on the available data.

1.5.3. General Approach to Classification Problem

A classification technique, or a classifier, is a systematic approach to building classification models from an input data set. Examples include, Decision Tree Classifiers, Rule-Based Classifiers, Neural Networks, Support Vector Machines and Naïve Bayes Classifiers.

Each technique employs a learning algorithm to identify a model that best fits the relationship between the attribute set and class label of the input data. The model generated by a learning algorithm should both fit the input data well and correctly predict the class labels of records it has never seen before. Therefore, a key objective of the learning algorithm is to build models with good generalization capability, i.e., models that accurately predict the class labels of previously

unknown records. Figure 1.14 shows a general approach for solving classification problems.

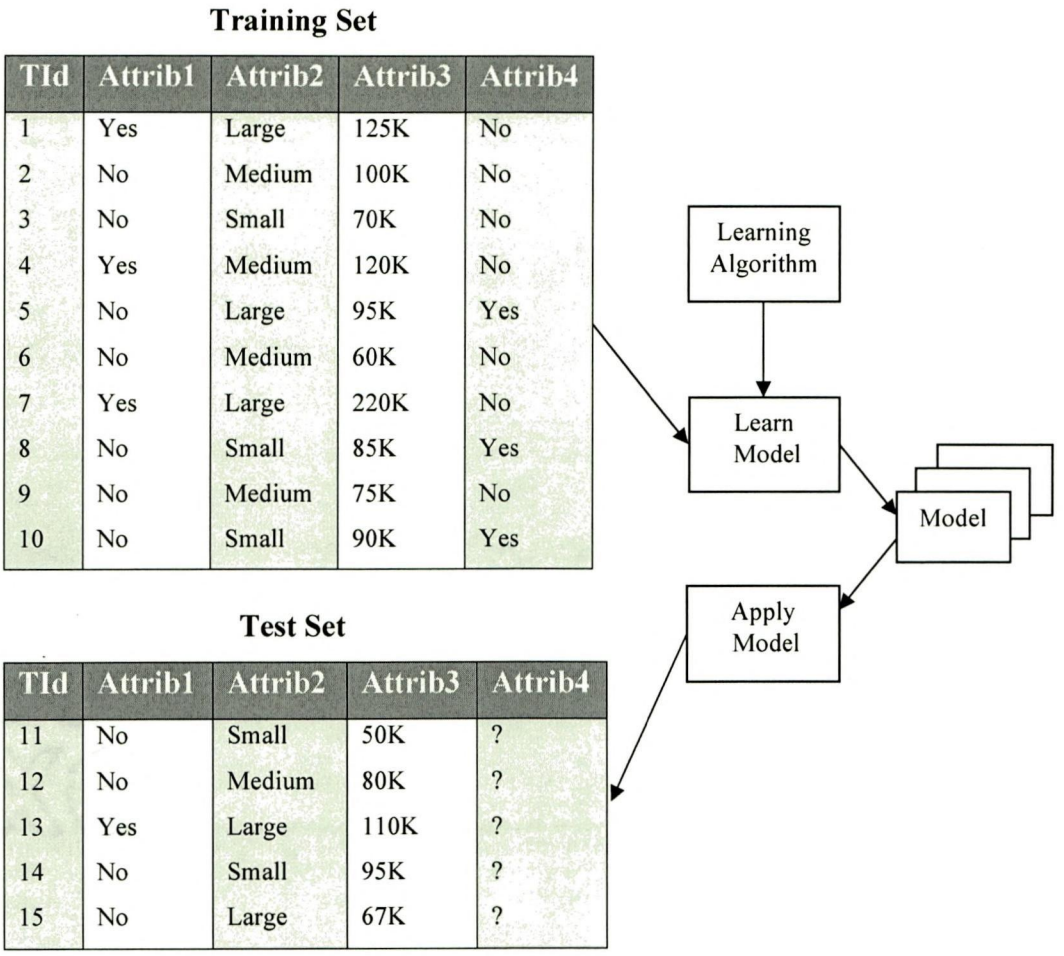

Figure 1.4: Process of Classification

First, a training set consisting of records whose class labels are known must be provided. The training set is used to build a classification model, which is subsequently applied to the test set, which consists of records with unknown class labels.

1.6. IMAGE CLASSIFICATION

Assigning images to pre-defined categories by analyzing the contents is defined as 'Image classification or 'Image categorization' (Chan et al., 2006). The process of image classification allows users to find desired information faster by searching only the relevant categories and not the whole information space. Image classification normally involves the processing of two main tasks.

- Feature extraction task – extract image features and forms feature vectors
- Classification task – uses the extracted features to discriminate the classes.

Three paradigms can be identified during the classification (Figure 1.5) and are listed below.

(i) Binary case

(ii) Multi-class case

(iii) Multi-label case

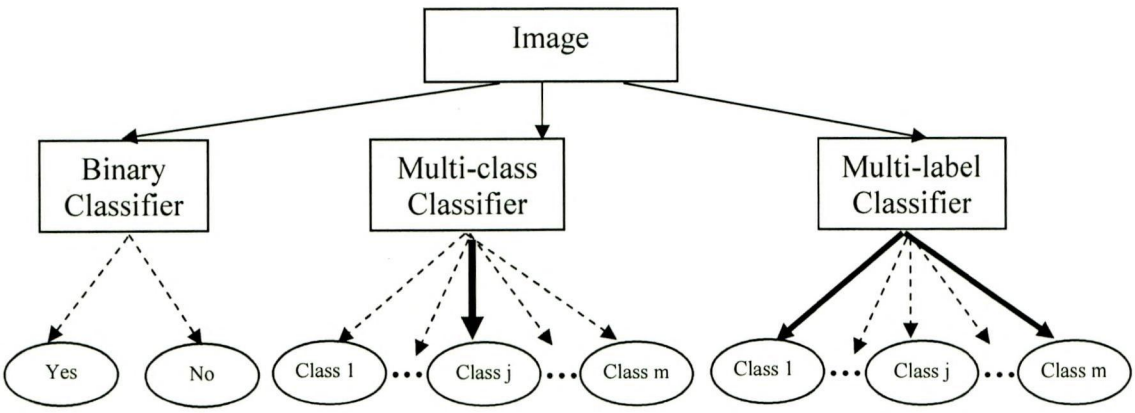

Figure 1.5: Paradigms in Image Classification

The binary case classification classifies images into exactly two predefined classes. Here, a sample image belongs exactly to one of the two given classes. The

classifier has to determine to which of the two sets the new image goes (Mehta *et al.*, 2008). In mutli-class case, an image belongs exactly to just one class of a set of 'm' classes (Foody and Mathur, 2004; Joshi *et al.,* 2009). Finally, in the multi-label case, an image may belong to several classes at the same time, that is, classes may overlap (Li *et al.*, 2004).

In binary classification a classifier is trained, by means of supervised algorithms, to assign a sample document to one of the two possible sets. These two sets are usually referred to as belonging samples (positive) and not belonging samples (negative) respectively. This method is otherwise termed as the one-against all approach or one-against one approach. Several algorithms exist for this type of classification. They are Naive Bayes, Linear Regression, Support Vector Machines (SVM) (Joachims, 1998) and LVQ (Martin-Valdivia *et al.*, 2003). The binary case has been set as a base case from which the other two cases, multi-class and multi-label, are built.

In multi-class and multi-label cases, the traditional approach consists on training a binary classifier for every class and then whenever the binary base case returns a measure of confidence on the classification, assigning either the top ranked one (multi-class assignment) or a given number of the top ranked ones (multi-label assignment). More details about these three paradigms can be found in (Allwein *et al.*, 2000). The proposed fusion-based image classifier defines multi-class classifiers.

1.7. MULTI-CLASS CLASSIFICATION

In machine learning, multiclass or multi-label classification is the special case within statistical classification of assigning one of several class labels to an input image. Unlike the well-understood problem of binary classification, the multiclass one is a more complex and less researched problem. Multiclass classification is often carried out by serially applying binary classification. This

can be accomplished by various strategies, including "One versus All" (OvA), "All versus All" (AvA) or more sophisticated information theoretic approaches. The OvA approach relies on the existence of a single computationally simple criterion which separates one class from the rest and makes use of standard binary classifiers. The AvA approach assumes the existence of simple separators between each pair of classes (Becker *et al.*, 2003).

In the past, multi-label classification was mainly motivated by the tasks of text categorization and medical diagnosis. Today, multi-label classification methods are seeing increased use in applications and some of them are discussed below.

- Classification of medical imagery (x-rays, MRI scans or digital photos of cells) to aid in the diagnosis of a variety of injuries and diseases.

- Analysis of maps to determine the proportion of various covers (asphalt, trees, soil, roofing) may be necessary. Classification of certain kinds of objects in images which are primarily "background" is also often necessary. An example of this kind of problem is the classification of ships in synthetic aperture radar imagery.

- Optical character recognition of digits and letters, either handwritten or typed, represent a third important category of multi-class image classification.

- A fourth important kind of image classification task is the recognition of a number of possible faces (people who need to be identified and monitored for security reasons) from a crowd. Such a situation may arise at an airport where cameras may programmatically scan faces, checking for the presence of wanted criminals.

- The classification of medical and biological data represents another important kind of multi-class classification task. An example of this kind is the classification of cells as either non-cancerous or as one of several types of cancer through examination of the gene expression profile.

- Classification of the structure of protein folds in various proteins.

- Another most frequently used type of multi-class classification problem is the classification of text documents. This is done to improve the semantic accuracy of document searches by reducing the number of false-positives.

1.8. CLASSIFIER COMBINATION

Broad classes of statistical classification algorithms have been developed and applied successfully to a wide range of real-world domains. In general, ensuring that the particular classification algorithm matches the properties of the data is crucial in providing results that meet the needs of the particular application domain. One way in which the impact of this algorithm/application match can be alleviated is by using group of classifiers, where a variety of classifiers (either different types of classifiers or different instantiations of the same classifier) are pooled before a final classification decision is made. Intuitively, fusion classification allows the different needs of a difficult problem to be handled by classifiers suited to those particular needs. Mathematically, fusion classifier provide an extra degree of freedom in the classical bias/variance tradeoff, allowing solutions that would be difficult (if not impossible) to reach with only a single classifier. Because of these advantages, fusion classification has been applied to many difficult real-world problems. A general model of fusion classification is presented in Figure 1.6.

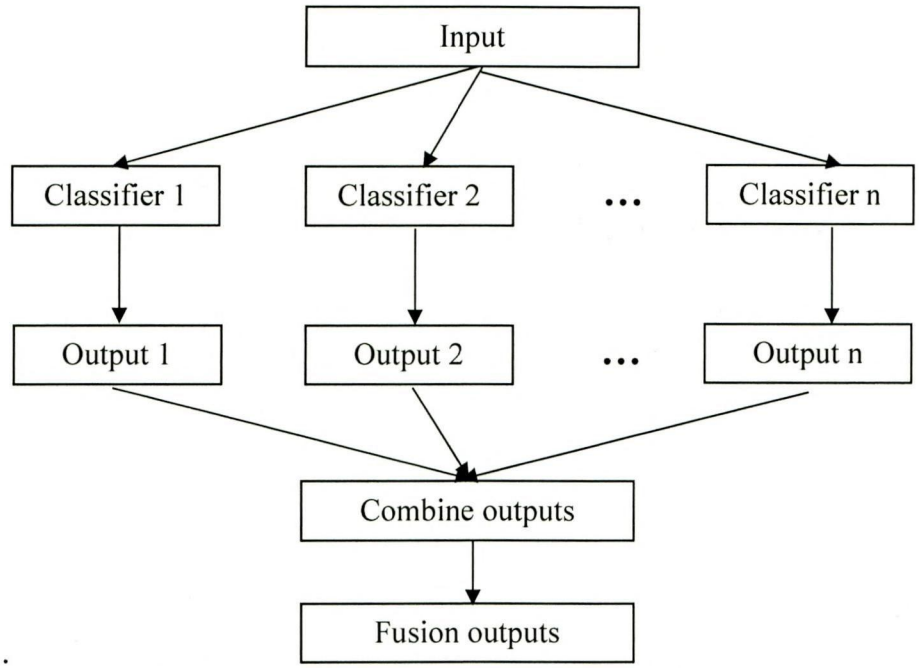

Figure 1.6: Classifier Combination Model

Recently, many scholars make use of classifier combination to enhance the performance of classification. In the past several years, a lot of effort has been devoted to different fusion methods to achieve better performance. In reality, how to select appropriate classification methods towards image classification is an unsolved problem (Oza and Tumer, 2008).

1.9. MOTIVATION

Large repositories of digitized books and manuscripts are emerging worldwide. Providing access to these collections require the conversion of these images to textual form with the help of Optical Character Recognizers (OCRs). Design of robust OCRs is still a challenging task for Indian scripts. The central module of an OCR is a recognizer which can generate a class label for an image component. Classification of isolated characters and thereby recognizing a complete document is still the fundamental problem in most of the Indian languages.

Characters are first segmented out from page or word images. A set of appropriate features are then extracted for representing the character image. Features could be structural or statistical. Structural features are often considered to be sensitive to degradations in the print. A feature-vector representation of the image is then classified with the help of a classifier. Classification of characters in Indian scripts is challenging due to

(i) Large number of classes

(ii) Many pairs of very similar characters.

A strong requirement of any robust character recognition system is the high classification accuracy.

According to Park *et al.* (2010) when a perfect set of features that can describe the image data is given, the accuracy of the resultant classification depends on the classifier adopted. Several solutions have been proposed for this purpose. Among which, the usage Neural Network (NN), K Nearest Neighbor (KNN) and Support Vector Machines (SVM) based classifiers are more prominent. The reasons behind this popularity are

(i) Easy of implementation procedures and

(ii) Accurate classification.

However, as pointed out by Neeba and Jawahar (2009) the success rate of character classification problem of characters in Indian scripts can be improved by using multiple classifiers. Motivated by this statement, the present research work combines classifiers to classify TIHC images into classes. The problem statement and objectives of the study is outlined in the following section.

1.10. PROBLEM STATEMENT AND RESEARCH OBJECTIVES

The problem statement of the present research work is formulated as given in Figure 1.7.

To develop a Classifier combination for the above problem statement, the following sub-objectives were framed.

- To develop efficient classification algorithms to classify TIHC characters using visual features.
- To develop and implement classifiers based on Neural Networks (NN), K Nearest Neighbor (KNN) and Support Vector Machine(SVM) that can efficiency classify images.
- To proposed 2-classifer and 3-classifer fusion algorithms that respectively combine any of the two and three of the single classifiers NN, KNN and SVM for TIHC image classification.
- To conduct a performance analysis to analyze the effectiveness of the proposed fusion classifiers in terms of accuracy and speed of classification and compare them against their single classifier counterparts.

"Let '

- *C_L' be a set of 'c' class labels, e.g, 'aa', 'kaa', 'kai' that determines the set of 256*** tamil characters.*
- *$F = \{f_1, f_2, \ldots, f_n\}$ be an n-dimensional feature column-vector describing the features of a Tamil character image. Each component 'f_i' of 'F' expresses features of a Tamil Character Image 'i' such as area, mean, standard deviation, minimum and maximum intensity and median.*

The classifier is any mapping

$$D : R_n \rightarrow [0, 1]c$$

i.e., the output D(x) is a c-dimensional vector whose ith component denotes the "support" for the hypothesis that comes from class. To further combine the results of a set of classifiers to produce the final classification decision. The combination is designed as an aggregation of the outputs of N individual classifiers, i.e.,

$$D(x) = AO(D_1), \ldots AO_N(x)$$

where AO is an aggregation operator. If a single class label is needed for F, the class with the highest support is chosen."

Figure 1.7: Problem Statement

1.11. LAYOUT OF THE THESIS

The underlying objective of this research work is to classify TIHC images using multiple classifiers and fuse their result for accurate classification. The working of the algorithms and the results obtained during experimentation are analyzed to determine the winner combination of classifiers reported in this dissertation. This chapter presented a brief outline to Tamil Scripts, OCR and

introduced the concept of image classification along with the problem statement and research objectives. The rest of the chapters are organized as follows.

The literature review is a critical look at the existing research that is significant to the work that is carried out. In case of character recognition several researchers have addressed the problem of character image classification. A critical look at the various available literatures related to the present research work is given in Chapter 2, Review of Literature.

The main components of the selected algorithm are the three selected classifiers, namely, Neural Networks, KNN and SVM. The working of these classifiers and the method of fusion is presented in Chapter 3, Methodology.

The proposed fusion classifiers were tested with several TIHC images for evaluating the best combination of classifiers for efficient classification. Chapter 4 tabulates and discusses the various results obtained.

The conclusion of the research work is summarized along with future research direction in Chapter 5. The work of several researchers are quoted and used as evidence to support the concepts explained in this dissertation. All such evidences used are listed in the reference section of the dissertation.

1.12. CONCLUSION

This chapter provided a brief introduction to the research problem, that is, classification of handwritten Tamil image scripts. The objectives formulated were also outlined. To achieve the objectives outlined in this chapter, a review of the previous research work was studied and the scrutinized works are summarized in the next chapter, Review of Literature.