"Statistical Techniques for Feature Extraction for Handwritten Character Recognition"- A Survey

P.E.Ajmire[1] Rajeev Dharaskar[2] V M Thakare[3]

1. Dept. Of Comp.Sci. , G. S. Sci., Arts & Comm. College, Khamgaon.
2. Dept. Of Comp. Sci., G H R C O E, Nagpur.
3. Dept. Of Computer Science, S G B Amravati University, Amravati.

Abstract

Keyword: Statistical techniques, Feature Extraction, Handwritten character recognition.

The content work in this paper is to describe the statistical techniques used for feature extraction for handwritten character recognition. Many researchers are working for recognition of handwritten characters. There are many script and languages in the world. The researchers have done work on some of them like English, Chinese, Latin, Arabic, Japanese, Thai, Urdu, Bangala, Telgu, Gurumukhi and Devnagari. The character set of Indian languages is large and consists of more complex characters when compared to the Latin script. Handwritten character recognition being a challenging problem in pattern matching area. There are various techniques used for this task like structural, neural and template matching. Since every electronic image of a character consist of pixel values that are represented by spatial configuration of O's and 1's. A statistical technique for character recognition is searching of statistical characteristics of various characters. The object of this study is to verify the applicability of these statistical techniques such as PCA, LDA, ICA, SVM to handwritten character recognition.

1. Introduction

Pattern recognition encompasses two fundamental tasks: description and classification. Given an object to analyze, a pattern recognition system first generates a description of it (i.e., the pattern) and then classifies the object based on that description (i.e., the recognition). During the last four decades, the field of character recognition has been receiving significant attention. Character recognition has been an active research point in pattern recognition field for several decades. Two general techniques for implementing pattern recognition systems, statistical and structural, employ different techniques for description and classification. Statistical techniques to pattern recognition use decision theoretic concepts to discriminate among objects belonging to different groups based upon their quantitative features. Among these techniques, statistical techniques are always favored in practical application due to their robust characteristics and simple training schemes. Many researchers from different disciplines are working in this field. Most of the work done in the field of character recognition is confined to Roman, English, Arabic [1], Japanese [2], Thai [3] and Devnagari [4], Chinese[5] .

Now a day some efforts have been reported in literature for Devnagari[6], Bangala[7], Tamil[8], Malayalam[9], Gurumukhi[10]. Although there are many scripts and languages in India but not much research work is done for handwritten Marathi characters[11]. Marathi handwritten character recognition is the challenging task in the pattern recognition field. Different statistical methods for Marathi handwritten character recognition have been proposed in recent years.

2. Feature Selection

The features of the character are the distinct characteristics of that character which help to recognize it. For pattern recognition the features should be easily computed and robust. Two types of features used in pattern recognition. The first one has clear physical meaning, such

as structural or statistical features. Another type of features has no physical meaning, but these features are treated as mapping features. The advantage of mapping features is that they can make classification easier.

2.1 Introduction to statistical pattern recognition

i. Formulation of the problem: gaining a clear understanding of the aims of the investigation and planning the remaining stages.
ii. Data collection: making measurements on appropriate variables and recording details of the data collection procedure (ground truth).
iii. Initial examination of the data: checking the data, calculating summary statistics and producing plots in order to get a feel for the structure.
iv. Feature selection or feature extraction: selecting variables from the measured set that are appropriate for the task. These new variables may be obtained by a linear or nonlinear transformation of the original set (feature extraction). To some extent, the division of feature extraction and classification is artificial.
v. Unsupervised pattern classification or clustering. This may be viewed as exploratory data analysis and it may provide a successful conclusion to a study. On the other hand, it may be a means of preprocessing the data for a supervised classification procedure.
vi. Apply discrimination or regression procedures as appropriate. The classifier is designed using a training set of exemplar patterns.
vii. Assessment of results. This may involve applying the trained classifier to an independent test set of labeled patterns.
viii. Interpretation.

In the statistical technique, each pattern is represented in terms of n features or measurements and is viewed as a point in a n-dimensional space. The goal is to choose those features that allow pattern vectors belonging to different categories to occupy compact and disjoint regions in a n-dimensional feature space. The effectiveness of the representation space (feature set) is determined by how well patterns from different classes can be separated. Given a set of training patterns from each class, the objective is to establish decision boundaries in the feature space which separate patterns belonging to different classes. In the statistical decision theoretic technique, the decision boundaries are determined by the probability distributions of the patterns belonging to each class, which must either be specified or learned. One can also take a discriminant analysis-based technique to classification: First a parametric form of the decision boundary (e.g., linear or quadratic) is specified; then the best decision boundary of the specified form is found based on the classification of training patterns. Such boundaries can be constructed using, for example, a mean squared error criterion.

3. Statistical Features

It is an technique to machine intelligence which is based on statistical modeling of data. With a statistical model in hand, one applies probability theory and decision theory to get an algorithm. Features are the measurements which represent the character such as size, shape and intensity. The statistical model one uses is crucially dependent on the choice of features. Hence it is useful to consider alternative representations of the same measurements (i.e. different features). For example, different representations of the character values in an image. General techniques for finding new representations include discriminant analysis, principal component analysis, and clustering.

4. Feature extraction techniques

4.1 Discriminant Analysis

Constructing new features via linear combination so that classification is easier. The notion of "easier" is quantified

by Fisher's criterion, which applies exactly to Gaussian classes with equal variance and approximately to other models. Variants like Flexible discriminant analysis consider nonlinear combinations as well. Recognition of character in quadratic classifier by using discriminant function[12].

4.2 Principal Component Analysis

Constructing new features which are the principal components of a data set. The principal components are random variables of maximal variance constructed from linear combinations of the input features. Equivalently, they are the projections onto the principal component axes, which are lines that minimize the average squared distance to each point in the data set. To ensure uniqueness, all of the principal component axes must be orthogonal. PCA is a maximum-likelihood technique for linear regression in the presence of Gaussian noise on both inputs and outputs. class-dependent principal component analysis (PCA) and linear discriminant analysis (LDA) which gives improved performance as compared with other standard techniques when experimented on several machine learning corpuses[13].

4.3 Principal curve

Principal curves are smooth curves that minimize the average squared orthogonal distance to each point in a data set. Fitting a principal curve is a maximum-likelihood technique for nonlinear regression in the presence of Gaussian noise on both inputs and outputs. Principal points are individual points that minimize the average distance to each point in a data set. To reduce the effect of the noise, the Gaussian filter[14] is used as a preprocessing of characters. (they are the output of k-means).

4.4 Factor analysis

A generalization of PCA is based explicitly on maximum-likelihood. Like PCA, each data point is assumed to arise from sampling a point in a subspace and then perturbing it with full-dimensional Gaussian noise. The difference is that factor analysis allows the noise to have an arbitrary diagonal covariance matrix, while PCA assumes the noise is spherical. In the paper of A biometric identity system Principal Component Analysis-Linear Discriminant Analysis Feature Extractor for Pattern Recognition is used by A. Khan, H. Farooq[15].

4.5 Independent Component Analysis

Constructing new features are the independent components of a data set. The independent components are random variables of minimum entropy constructed from linear combinations of the input features. The entropy is normalized by the variance of the component, so absolute scale doesn't matter. It is a fact of information theory that such variables will be as independent as possible. This feature extraction technique is closely related to exploratory projection pursuit, commonly used for visualization.

4.6 Clustering

Grouping similar objects in a multidimensional space is useful for constructing new features which are abstractions of the existing features. Some algorithms, like k-means, simply partition the feature space. The quality of the clustering depends crucially on the distance metric in the space. Most techniques are very sensitive to irrelevant features, so they should be combined with feature selection. A shape feature computed from certain directional view base strokes of input character image, has been used by both HMM and ANN classifier of the Devnagari numeral recognition system[16].

4.7 K-means

A parametric algorithm for clustering data into exactly k clusters. First, define some initial cluster parameters. Second, assign data points to clusters. Third, modelling better cluster parameters, given the data assignment. HMM is a popular model for handwriting recognition because of its effectiveness and robustness. The features of the images of characters, significantly boosting the effectiveness of k-means clustering[17].

5. Statistical Classification Techniques.

The statistical technique has been most intensively studied and used in practice. More recently, neural network techniques and methods imported from

statistical learning theory have been receiving increasing attention. The design of a recognition system requires careful attention to the following issues: definition of pattern classes, sensing environment, pattern representation, feature extraction and selection, cluster analysis, classifier design and learning, selection of training and test samples, and performance evaluation.

Statistical classifiers are rooted in the baye's decision rule and can be divided into parametric ones and non parametric ones. Parametric classifier includes the linear discriminant function (LDF), the quadratic discriminant function (QDF), the Gaussian mixture classifier, etc. An improvement to QDF, named regularized discriminant analysis (RDA), was shown to be effective to overcome inadequate sample size and stabilized the performance of QDF by smoothing the covariance matrices. The modified quadratic discriminant function (MQDF) proposed by Kimura et al. was shown to improve the accuracy, memory, and computation, efficiency of the QDF. The directions are sampled down using Gaussian filter to get 392 dimensional feature-vectors. This feature vector is applied on MQDF classifier. The invariant moments are well known to be invariant under translation, scaling, rotation and reflection [18]. They are measures of the pixel distribution around the centre of gravity of the character and allow to capture the global character shape information. Traditionally, moment invariants are computed based on the information provided by both the shape boundary and its interior region. The moments used to construct the moment invariants are defined in the continuous but for practical implementation they are computed in the discrete form.

6. Current trends and outlook for the future.

There are many script and Languages in the world. The researchers have done work on some of them like English, Chinese, Latin, Arabic, Japanese, Thai and Devnagari. India is a multi-lingual and multi-script country comprising of eleven different scripts and not much work has been done towards handwritten recognition of Indian scripts. Recognition of handwritten characters is important because of its applicability to a number of problems, like postal code recognition and information extraction from fields of different forms. In the Indian context, there exists a need for development and/or evaluation of the existing techniques for recognition of handwritten character in Indian scripts. A variety of statistical techniques, model and techniques has emerged, influenced by developments in the field of pattern recognition such as character recognition, face recognition, finger print recognition, etc. In this paper, we have mentioned the statistical features extraction techniques and statistical techniques for its recognition. Selection of a feature extraction method is probably the single most important factor in achieving high recognition performance in character recognition systems. Different feature extraction methods are designed for different representations of the characters, such as solid binary characters, character contours, skeletons (thinned characters), or gray level sub-images of each individual character. The feature extraction methods are discussed in terms of invariance properties, re-constructability, and expected distortions and variability of the characters. When a few promising feature extraction methods have been identified, they need to be evaluated experimentally to find the best method for the given application.

7. Discussion

These statistical methods give the recognition rate between 85% to 90%. This recognition rate is increased by using some hybrid methods such as structural and statistical methods. Current research employs models not only of characters, but also words and phrases, and even entire documents, and powerful tools such as HMM, neural nets, contextual methods are being brought to bear. It is hoped that this comprehensive discussion will provide insight into the concepts involved, and perhaps provoke further advances in this area.

7. References :

1. I.A. Jannoud "Automatic Arabic Hand Written Text Recognition System" American Journal of Applied Sciences 4 (11): 857-864, 2007.
2. Tour Wakahara, Y. Kimura & Mutsuo " Handwritten Japanees Character Recognition Using Adaptive Normalization by Global Affine Transformation." Proc. 6th ICDAR Vol., Issue , 2001.
3. J.L.Mitrpanont, U. Limkonglap "Using Countour Analysis to improve Feature Extraction in Thai Handwritten Character Recognition Systems" Proc. 7th IEEE ICCIT , 2007.
4. U.Pal, N.Sharma, T.Wakabayashi and F. Kimura. "Off-line Handwritten character recognition of Devnagari Script" 9th ICDAR, 2007.
5. Bing Feng, Xiaoqing Ding, Ypushou Wu "Chinese Handwriting Recognition using Hidden Markov Models" p no 1051-4651/02 IEEE 2002.
6. H. Swethalakshmi1, Anitha Jayaraman1, V. Srinivasa Chakravarthy2, C. Chandra Sekhar "Online Handwritten Character Recognition of Devanagari and Telugu Characters using Support Vector Machines" Indian Institute of Technology Madras, Chennai - 600 036, India.
7. T.K.Bhowmik, A Roy & U Roy "Character Segmentation for Handwritten Bangla Words using Artificial Neural Network" Proc.1st IAPR TC3 NNLDAR, 2005.
8. R.Indra Gandhi, K Iyakutti, "An attempt to recognize Handwritten Tamil Character using Kohonen SOM" 188-192 Vol 1(3) IJANA 2009.
9. Bindu S Moni, G RAju "Modified Quadratic Classifier and directional features for Malayalam Character recognition", NCCSE 2011.
10. Kartar Singh Siddharth, Renu Dhir, Rajneesh Rani, "Handwritten Gurumukhi Character Recognition Using Zoning Density and Background Directional Distribution Features" ,1036-1041 Vol. 2 (3) , IJCSIT 2011.
11. P E Ajmire, R V Dharaskar and V M Thakare " A Comparative study of Handwritten Marathi Character Recognition", pp 29-32, NCIPET 2012
12. U.Pal, N.Sharama, T. Wakabayashi and F.Kimura "offline handwritten character recognition of Devnagari script" , ICDAR, IEEE 2007.
13. Alok Sharma, Kuldip K. Paliwal, Godfrey C. Onwubolu "Class-dependent PCA, MDC and LDA :A combined classifier for pattern classification Pattern Recognition" Society, Elsevier Ltd. 2006
14. N. Joshi, G. Sita, A.G. Ramakrishnana, Deepu V, S.Mahdavnath"Machine recognition of online handwritten Devnagari characters" .
15. A. Khan1, H. Farooq " Principal Component Analysis-Linear Discriminant Analysis Feature Extractor for Pattern Recognition", IJCSI Issues, Vol. 8, 2011
16. U. Bhattacharya, S.k. .Parui, B. Shaw, K. Bhattacharya "Neural Combination of ANN and HMM for handwritten Devanagari Numeral Recognition" .
17. Weijie Su and Xin Jin "Hidden Markov Model with Parameter-Optimized K-Means Clustering for Handwriting Recognition".
18. P.E.Ajmire and S E Warkhede "Handwritten Marathi Character (Vowel) Recognition" Advances in information mining, Vol 2, 2010.