Available online at www.elixirpublishers.com (Elixir International Journal)

Electrical Engineering

Elixir Elec. Engg. 38 (2011) 4390-4393

ISSN: 2229-712X

Statistical classifier with barcode based feature vectors for numerals recognition

Shreedharamurthy S K[1] and H.R.Sudarshana Reddy[2]

[1]Department of E&C, UBDT College of Engineering, Davangere-577004, Karnataka-India

[2]Department of E&E, UBDT College of Engineering, Davangere-577004, Karnataka-India.

ARTICLE INFO

Article history:

Received: 22 August 2011;

Received in revised form:

26 August 2011;

Accepted: 31 August 2011;

Keywords

Pattern recognition,

Feature extraction,

Statistical classifier,

Numeral recognition.

ABSTRACT

Selection of feature extraction method is most important factor in achieving high recognition performance in automatic pattern recognition systems. Similarly selection of suitable classifier also plays a very important role for the same. Plenty of feature selection methods and classifiers are existing in computer domain and choice of each of these mainly depends on task in hand. This paper presents an efficient and novel method for recognition of handwritten numerals using bar codes. Handwritten numerals are scan converted to binary images and normalized to a size of 30 x 30 pixels. The features are extracted using barcodes and are classified successfully using the statistical technique.

Introduction

Pattern recognition is the assignment of some sort of output value to a given input value (or instance), according to some specific algorithm. An example of pattern recognition is classification, which attempts to assign each input value to one of a given set of classes.

Pattern recognition is generally categorized according to the type of learning procedure used to generate the output value. Supervised learning assumes that a set of training data (the training set) has been provided, consisting of a set of instances that have been properly labeled by hand with the correct output. A learning procedure then generates a model that attempts to meet two sometimes conflicting objectives: Perform as well as possible on the training data, and generalize as well as possible to new data. Unsupervised learning, on the other hand, assumes training data that has not been hand-labeled, and attempts to find inherent patterns in the data that can then be used to determine the correct output value for new data instances. A combination of the two that has recently been explored is semi supervised learning, which uses a combination of labeled and unlabeled data (typically a small set of labeled data combined with a large amount of unlabeled data). Note that in cases of unsupervised learning, there may be no training data at all to speak of; in other words, the data to be labeled is the training data.

Typical applications of pattern recognition are automatic speech recognition, classification of text into several categories, the automatic recognition of handwritten postal codes on postal envelopes etc.,[2]

The recognition of handwritten numerals has been the subject of much attention in pattern recognition because of its number of applications such as bank check processing, interpretation of ID numbers, vehicle registration numbers and pin codes for mail sorting. Promising feature extraction methods have been identified in the literature for recognition of characters and numerals of many different scripts. These include template matching, projection histograms, geometric moments, Zernike moments, contour profile, Fourier descriptors, and unitary transforms. A brief review of these feature extraction methods is found in [1]. Various methods have been proposed, and high recognition rates are reported, for the recognition of English handwritten digits [8 - 10]. The task of classification is to partition the feature space into regions corresponding to source classes or assign class confidences to each location in the feature space. Statistical techniques, neural networks, and more recently support vector machine (SVM) have been widely used for classification due to the implementation efficiency [1-5].

Numerals Recognition System

A typical pattern recognition system consists of four stage processes as follows:

Data acquisition

Preprocessing

Feature extraction

Classification

Data acquisition

In the data acquisition stage we capture the pattern to be classified and the same is given to the next stage for further processing.

Pre-processing

Pre-processing involves normalizing the raw data given to the computer so that the further processing is easier. In the initial stage of the process, the preprocessing steps, includes general signal processing algorithms and also more application-specific algorithms. This is a step where certain normalizations are done. The typical preprocessing operations involves noise reduction, size normalization, slant estimation and correction, thinning, segmentation etc.,

Feature Extraction

When the input data to an algorithm is too large to be processed and it is suspected to be redundant then the input data will be transformed into a reduced representation set of features. Transforming the input data into the set of features is called features extraction. If the features extracted are carefully chosen

Tele: +919449390926, +919448728576

E-mail addresses: sks_murthy@yahoo.com, hrsreddy@hotmail.com

it is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input. The set of features that are used makes up a feature vector, which represents each member of the population. Then, character recognition system classifies each member of the population on the basis of information contained in the feature vector [6]. Feature extraction algorithms attempt to reduce a large-dimensionality feature vector into a smaller-dimensionality vector that is easier to work with and encodes less redundancy.

Selection of a feature extraction method is probably the single most important factor in achieving high recognition performance in character recognition systems.

Bar code method used to extract features of numerals in our system.

Barcode Method

Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately. When performing analysis of complex data, one of the major problems stems from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computation power or a classification algorithm which overfits the training sample and generalizes poorly to new samples. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy. Best results are achieved when an expert constructs a set of application-dependent features.

Extraction of potential feature is an important component of any recognition system. Selection of potential features is probably the single most important factor in achieving high recognition performance. In this paper, bar code based features are considered as the potential features and used for the recognition of handwritten numerals. In this method the captured image (hand written numeral) first converted to bmp format and after preprocessing of the same is conducted, its features are extracted using barcode method.

The bmp format of an image contains only zeroes and ones as shown in fig 1.below:

00 0 1 1 0 0 0
0 0 1 0 1 0 0 0
0 0 0 0 1 0 0 0
0 0 0 0 1 0 0 0
0 0 0 0 1 0 0 0
0 0 0 0 1 0 0 0
0 0 0 0 1 0 0 0
0 0 1 1 1 1 1 0

Fig 1. BMP Image values

The features from the bmp image can be had considering the horizontal bars, vertical bars, diagonal bars etc., as shown in shaded areas of the following figures. At least ten features are to be obtained in order to have good recognition rate. In the meantime number of features may not be too many as it leads to complexity in further processing.

The features such as HF1, HF2, HF3, VF1, VF2, VF3, TD1, TD2, TD3, and TD4 of fig 2. are calculated using following algorithm:

1. Read the pattern in bmp format.
2. Choose suitable number of bars out of whole pattern
3. For each bar area, count number of'ON'cells and total number of cells present in that bar.
4. Feature value F is sum of values of 'ON' cells / total number of cells present in that bar.
5. Repeat this for all the chosen bars in the patterns.

Optimum number of features at optimum places of pattern gives us high rate of recognition as such. In numerals recognition system we need to discriminate between '8' and '3', '5' and '6', '7'and'1'etc.,

It has been argued that since feature selection is typically done in an off-line manner, the execution time of particular algorithm is not as critical as the optimality of feature subset it generates. While this is true for feature sets of moderate size, several recent applications particularly those in data mining and document classification, involve thousands of features. In such cases, the computational requirement of a feature selection algorithm is extremely important. As the number of facture subset evaluations may easily become prohibitive for large feature sizes, a number of suboptimal selection techniques have been proposed which essentially tradeoff the optimality of the selected subset for computational efficiency.[7]

Classification

Numerous classifiers are available in computer domain to classify the pattern to different classes such as statistical classifiers, structural classifiers, support vector machines (SVM) hidden markov model (HMM), artificial neural networks etc., In this paper we proposed a statistical technique to classify the handwritten numerals from which we obtained considerably good recognition rate.

Proposed statistical technique:

Let the features obtained for a given pattern are *f1, f2, f3,…….fn*. Assign them to a variable *a01*. Hence,

a01= [f1, f2, f3, ……fn];

Obtain the features for similar pattern of same class with different style of writing and store them in another variable *a02*. Repeat the procedure for as many writing styles are possible and assign the feature to *a03, a04,……..a0n.*

So we have*a01, a02, a03,……an* variables each with *f1, f2, f3,…….fn* features for the patterns of same class

Similarly we have *a11, a12, a13, …. a1n* variables with same set of features for the patterns of other class.

This process is repeated for all the patterns of all the classes.

So we have *a01, a02, a03, ……. a0n* variables for patterns of class-1

*a11, a12, a13 …… a1n*variables for patterns of class-2

----------an1, an2, an3 ……..ann variables for patterns of class-n.

Get the features f1, f2, f3, …… fn for the pattern to be classified and assign them to a variable a. Then get the difference between variable a and all other variables in database as follows:

a01=a-a01;a02=a-a02;a03=a-a03

…….a0n=a-a0n

a11=a-a11;a12=a-a12; a13=a-a13

…….a1n=a-a1n

---

an1=a-an1; an2=a-an2; an3=a-an3

…..ann=a-ann

We then compute the sum of absolute values for each class as

c01=∑abs(a01);c02=∑abs(a02);……..

……c0n=∑abs (a0n);class-1

c11=∑abs(a11);c02=∑abs(a12);……
…….……c 1n=∑abs(a1n);class-2

---
cn1==∑abs(an1);c02=∑abs(an2);……
…….……cnn=∑abs (ann);class-n
Among all these values c01, c02, c03 ……..……cnn we need find minimum value as
m = ;
Then the given pattern belongs to the class to which the value m belongs.

Results and Conclusion

The proposed feature extraction method and statistical technique to classify the hand written numerals was implemented using MATLAB-7 successfully. We used NIST database numerals. Few NIST database samples are as given in figure 3. Among thousands of samples, we used around 100 samples for each numeral out of which 80 samples used for creating statistical database and 20 samples are used for testing. We got 95% recognition if new pattern selected among 20 samples is used for classification. 100% recognition was obtained for the patterns selected among 80 samples used for creating statistical database.

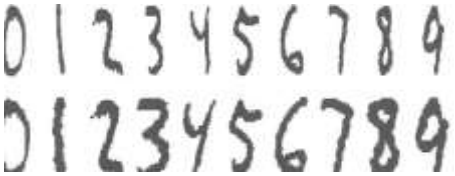


Fig 3. Samples of handwritten numerals

In this paper we proposed and implemented barcode based feature vectors along with the statistical method to classify the handwritten numerals. We achieved considerably good recognition.

References

[1] Øivind Due Trier, Anil K. Jain and TorfinnTaxt, Feature Extraction Methods for Character Recognition- A survey, Pattern Recognition, Volume 29, Issue 4, April 1996, pp641-662.

[2] K V Prema_ and N V Subbareddy- Two-tierarchitecture for unconstrained handwritten character recognition-Sadhana Vol. 27, Part 5,October 2002, pp.585-594.

[3] Hyun-Chul Kim, Daijin Kim, Sung YangBang-A numeral character recognition using PCA mixture model, pattern recognition letters,Vol 23, 2002, pp.103-111

[4] G Y Chen, T D Bui, A. Krzyzak-Contour based numeral recognition using muti wavelets and neural networks, pattern recognition letters,Vol 36, 2003, pp.1597-1604

[5] RejeanPlamondon, Sargur.N.Srihari, On-lineand Off-line Handwriting Recognition: A Comprehensive survey, IEEE Trans, Pattern Analysis and Machine Intelligence, vol 22, no 1,pp 63-79,Jan 2000

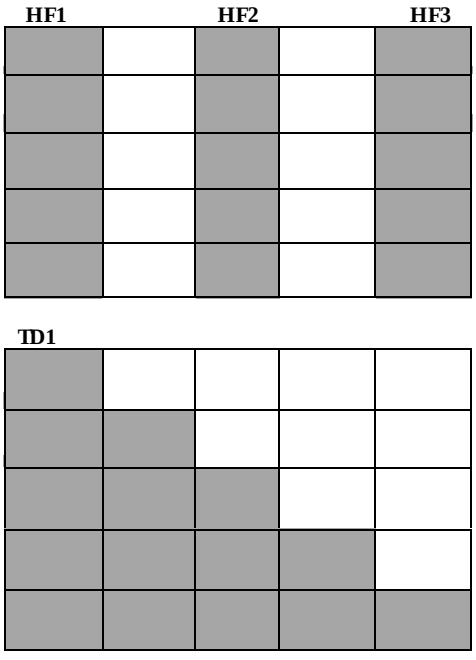
[6] Claus Bahlmann-Directional features inonline handwriting recognition, patternrecognition letters, Vol 39, 2006, pp.115-125

[7] Anil K Jain, Robert P W Dudin, Jianchang Mao-Statistical Pattern recognition: A Review, IEEE transaction on pattern analysis and machine intelligence vol 22, No 1, January, 2000.

[8] Alexander Goltsev, Dmitri Rachkovskij-Combination of the assembly neural network with a perceptron for recognition of handwritten digits arranged in numeral strings, pattern recognition letters, Vol 38, 2005, pp.315-322

[9] Bailing Zhang, Minyue Fu, Hong Yan-Anonlinear neural network model of mixture of local principal component analysis: application to handwritten digits recognition, pattern recognition letters, Vol 34, 2001, pp.203-214

[10] Cheng-Lin Liu, Kazuki Nakashima, HiroshiSako, Hiromichi Fujisawa, Handwritten digit recognition: investigation of normalization and feature extraction techniques, Jun 2003



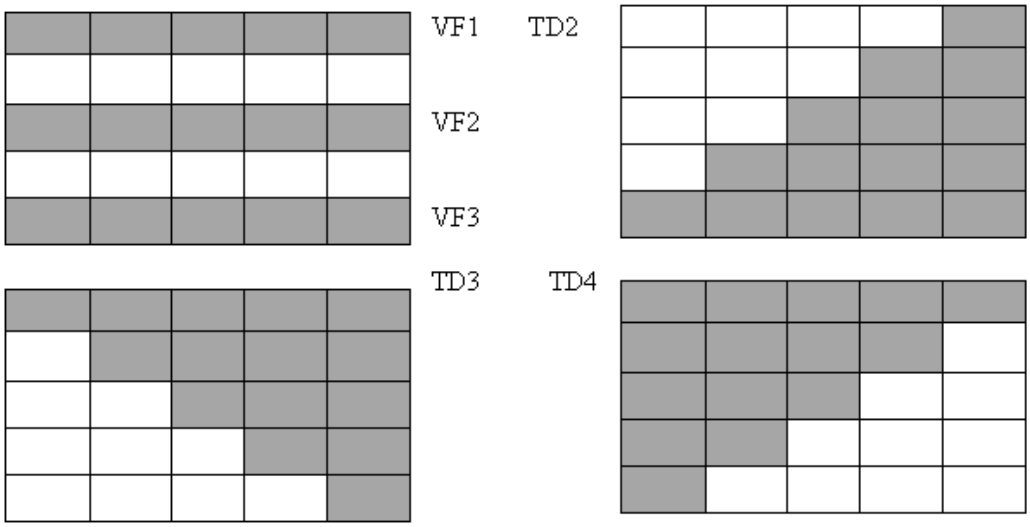


Fig 2: Pattern's future areas