Handwritten Character Recognition System using Chain code and Correlation Coefficient

Ravi Sheth1, N C Chauhan2, Mahesh M Goyani3, Kinjal A Mehta4

1Information Technology Dept., A.D Patel Institute of Technology, New V V nagar-388120, Gujarat, India
2Information Technology Dept., A.D.Patel Institute of Technology, New V V nagar-388121, Gujarat, India
3Computer Engineering. Dept., L.D.college of engineering, Ahmadabad, Gujarat, India
4Electronics and Communication Dept., L.D. college of engineering, Ahmadabad, Gujarat, India

ABSTRACT— Pattern recognition deals with categorization of input data into one of the given classes based on extraction of features. Handwritten Character Recognition (HCR) is one of the well-known applications of pattern recognition. For any recognition system, an important part is feature extraction. A proper feature extraction method can increase the recognition ratio. In this paper, a chain code based feature extraction method is investigated for developing HCR system. Chain code is working based on 4-neighborhood or 8–neighborhood methods. In this paper, 8–neighborhood method has been implemented which allows generation of eight different codes for each character. These codes have been used as features of the character image, which have been later on used for training and testing for Neural Network (NN) and Support Vector Machine (SVM) classifiers. In this work we have also implemented HCR system with the use of correlation coefficient. Comparison of all the methods for HCR systems are highlighted at the end.

Keywords: Pattern recognition, handwritten character recognition, feature extraction, chain code, correlation coefficient, neural network, support vector machine.

1.INTRODUCTION

pattern recognition is a field of study whose general goal is the classification of objects into a number of categories. The first process of this mechanism is to design a dataset for feature extraction and another data set is for to train the classifier. In next process different feature extraction methods are applied on input data set and extract feature from it. This extracted feature applied on different classifier which was already trained trough input data set is recognized pattern based on match between feature and trained data. Pattern recognition system can be used in so many applications such as face recognition, character recognition, and speech recognition. It has been lots of work is carried out in each and every application. In this paper we have presented handwritten character recognition technique.

A HCR is one of the highly used applications of pattern recognition techniques. Handwriting recognition has always been a challenging task in pattern recognition. Many systems and classification algorithms have been investigated by various researchers since past few decades. Techniques ranging from mathematical methods such as principle component analysis and fisher discriminate analysis [1] to machine learning like artificial neural networks [2] or support vector machines [3] have been used to solve this problem. Generally, the steps of HCR can be divided in four major parts as shown in Fig. 2. These phases include binarization, segmentation, feature extraction and classification [4]. Binarization refers to the conversion of a gray-scale image into binary images using the appropriate threshold [4]. After binarization next step is segmentation. Due to the wide range of changes available into handwritten characters, it is very difficult to recognize the characters correctly.

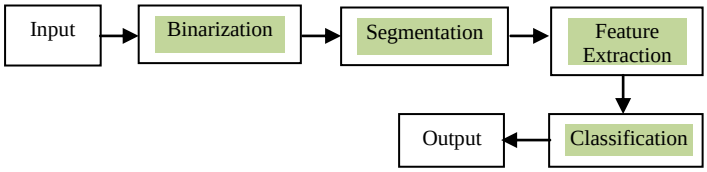


Figure 1: Block diagram of HCR system.

But before recognition, the handwritten characters have to be processed to make them suitable for recognition. Here, we consider the processing of entire document containing multiple lines and many characters in each line. Our aim is to recognize characters from the entire document. The handwritten document has to be free from noise, skew ness, etc. The lines and words have to be segmented. The characters of any word have to be free from any slant angle so that the characters can be separated for recognition. By this assumption, we try to avoid a more difficult case of cursive writing. Segmentation of unconstrained handwritten text line is difficult because of inter-line distance variability, base-line skew variability, different font size and age of document [4]. During the next step of this process features are extracted from the segmented character. Feature extraction is a very important part in character recognition process. Extracted feature has been applied to classifiers which recognized character based on trained features. In second section, we have described feature extraction method in brief and described chain code methodology with 4 and 8 neighborhoods' method.

2.FEATURE EXTRCTION

Any given image can be decomposed into several features. The term 'feature' refers to similar characteristics. Therefore, the main objective of a feature extraction technique is to accurately retrieve these features. The term "feature extraction" can thus be taken to include a very broad range of techniques and processes to the generation, update and maintenance of discrete feature objects or images [5]. Feature extraction is the most difficult part in HCR system.

For feature extraction, programmers must manually determine the properties which they feel are important.

Some example properties [6] might be:

Aspect ratio(ratio of width of image to its height)
Percent of pixels above horizontal half point
Percent of pixels to right of vertical half point

Number of strokes
distance from image center
reflected y axis
reflected x axis
This approach gives the recognizer more control over the properties used in identification. Character classification task recognizes the character which is compared with the standard value that comes out the learning character, and the character should be corresponded to the document image that is matching a setting document style in the document style setting part.

CHAIN CODE GENERATION

The chain code extraction algorithm of HCR is shown in Fig. 2. Basically chain code is working based on two different manners such as 4- neighborhood method or 8-neighborhood method. In this work, we have implemented 8-neighborhood method for chain code. This method has been implemented as feature for English handwritten characters classification. In order to obtain the chain ode, we just focus on the main part (body) of the character image. From the first pixel of the image, we move downwards row by row and consider the first pixel of the body of image which exactly has got one neighbor, as the start point of the chain code [7]. If a character has no initial point, we will consider its chain code as zero. Each pixel of image has received its eight neighbors; to each neighbor we assign one value between 0 and 7. After finding the start point of the chain code neighbors of a pixel and value is assigned to it in a given character image, we move to the next neighbor pixel which also be a part of image. Again in the cases of having two or more neighbor pixels with the above condition use the directional priority. While passing one pixel to its next pixel, we have inserted the number related to that next in the chain code of that image [7]. After receiving the chain code for all segmented characters, we come to know that the chain code for different characters has different size, and size of each chain code depends on the length of the desired character. In more ever, length of the chain codes is usually high; therefore one should convert it into its normalize form. This chain code as its length will be fixed and limited [8].

Algorithm for generating chain code considering 8-neighborhood is as follows:

Input image

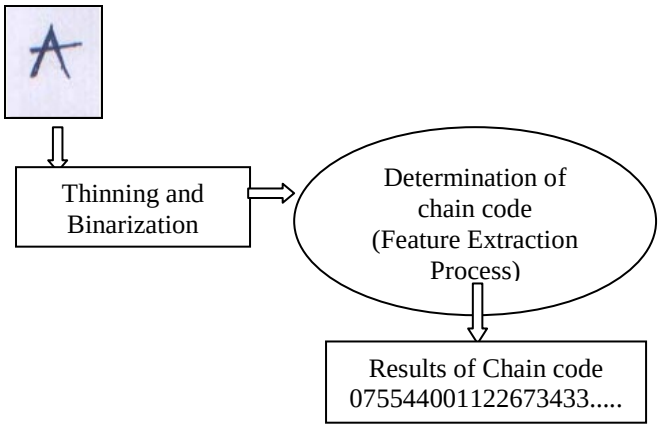


Figure 2: Chain code extraction for HCR

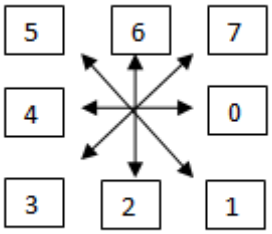


Figure 3: 8-neighbourhood for chain code

Step1: Find out starting point which has nonzero values and store it in first
Step2: Initialize 0-7 total eight directions
Step3: Travels all 8 neighbors
Step4: Find first nonzero value
Step5: Add it in to chain code list
Step6: Move to next position
Step7: Check whether we reach to first point or not if not then go to step 3.

Chain code normalization

After applying chain code on segmented image, we came to know that for any character length of chain code vary for different character. So in this case it is very difficult to keep all the chain code for all the characters. For example, we have applied chain code for figure 3 and result is shown in figure 6. From figure 5, we can see that for only one character we have to store 58 values. In order to normalize the obtained chain code, we transform it to a two dimensional matrix where in the first row, the value of the chain code, and in the second row, frequency of occurrence of that value are written [9]. This frequency of occurrences of different neighbours can also be considered as histogram of neighbouring indexes. The resultant normalized chain code of figure 5 is shown in figure 6. It should be noted that even though we obtain variable number of neighbourhood indices, due to histogram consideration, the number of features for each character becomes fixed which 8 for 8-neighbourhood method is.

Figure 4: Input image

0 1 1 2 2 2 2 2 1 2 1 2 1 2 2
1 2 2 1 2 1 2 1 2 1 2 1
0 1 5 4 5 6 5 6 5 6 5 6 5 6 6
5 6 6 5 6 5 6 5 6 6 6 6
6 5 5 4

Figure5: Chain code of figure 5

Normalized chain code:

2 12 15 0 2 12 15 0

Figure 6: Normalized chain code

C. Recognition using Correlation coefficient

The correlation coefficient is one of the popular metric used in the literature to provide comparison of two images. And well it should be, for our empirical knowledge is fundamentally of co-varying things. We come to discern relationships among things in terms of whether they change together or separately; we come to impute causes on the basis of phenomena co-occurring; we come to classify as a result of independent variation [10]. In this paper for recognition purpose, we have used correlation coefficient method [10]. The correlation coefficient given as,

$$R = \frac{\sum_m \sum_n \left(A_{mn} - \overline{A}\right)\left(B_{mn} - \overline{B}\right)}{\sqrt{\left(\sum_m \sum_n \left(A_{mn} - \overline{A}\right)^2\right)\left(\sum_m \sum_n \left(B_{mn} - \overline{B}\right)^2\right)}}$$

Where R indicates the Correlation Coefficient (CC) between two matrices A and B, where A and B are matrices or vectors of the same size. $\overline{A}$ = mean (A), and $\overline{B}$ = mean (B). A correlation coefficient is a numerical, descriptive measure of the strength of the linear relationship between two variables. Values for the correlation coefficient range between -1 and +1; with a correlation coefficient of +1 indicating that the two variables have a perfect, upward-sloping (+) linear relationship and a correlation coefficient of -1 showing that the two variables are perfectly related in a downward-sloping, (-) linear sense. A correlation coefficient of 0 demonstrates that the variables have no relationship, and are independent. In this paper, character recognition is carried out without the use of standard classifiers such as neural network, support vector machines, etc. With the use of Matlab tool templates are created and stored for all A to Z characters and for 0 to 9 digits in fixed size. This database is of optical characters data set. The templates are generated by averaging few database images of characters and digits. .After applying segmentation algorithm, we obtain segmented characters, which are resized to the dimension of the template characters. After resizing, cross correlation metric is found between segmented characters with all stored template images. The maximum value of CC metric indicate best match of the extracted character with the stored database character. The process is repeated for all the characters and finally the output is stored in text file.

classification methods

Neural Network

Artificial neural networks (ANN) provide the powerful simulation of the information processing and widely used in patter recognition application. The most commonly used neural network is a multilayer feed forward network which focus an input layer of nodes onto output layer through a number of hidden layers. In such networks, a back propagation algorithm is usually used as training algorithm for adjusting weights [9]. The back propagation model or multi-layer perceptron is a neural network that utilizes a supervised learning technique. Typically there are one or more layers of hidden nodes between the input and output nodes. Besides, a single network can be trained to reproduce all the visual parameters as well as many networks can be trained so that each network estimates a single visual parameter. Many parameters, such as training data, transfer function, topology, learning algorithm, weights and others can be controlled in the neural network [9].

Support Vector Machine

The main purpose of any machine learning technique is to achieve best generalization performance, given a specific amount of time and finite amount training data, by striking a balance between the goodness of fit attained on a given training dataset and the ability of the machine to achieve error-free recognition on other datasets [11].

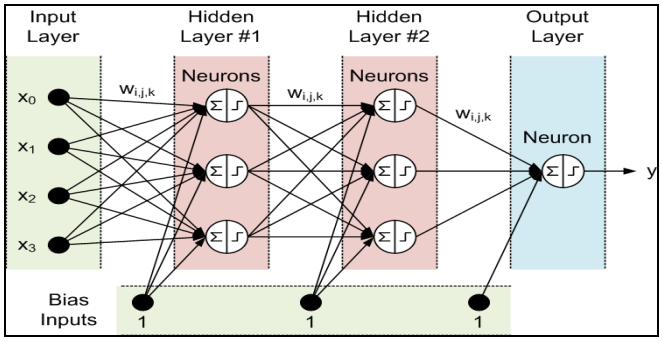


Figure 7: neural network design

With this concept as the basis, support vector machines have proved to achieve good generalization performance with no prior knowledge of the data. The main goal of an SVM [10] is to map the input data onto a higher dimensional feature space nonlinearly related to the input space and determine a separating hyper plane with maximum margin between the two classes in the feature space.

Main task of SVM is to finds this hyper plane using support vectors ("essential" training tuples) and margins (defined by the support vectors).

Let data D be (Z1, y1), …, (Z|D|, y|D|), where Xi is the set of training tuples associated with the class labels yi which has either +1 or -1 value [11].

There are uncountable (infinite) lines (hyper planes) separating the two classes but we want to find the best one (the one that minimizes classification error on unseen data). SVM searches for the hyper plane with the largest margin, i.e., Maximum Marginal Hyper plane (MMH) [11]. The basic concept of SVM can be summarized as,

A separating hyper plane can be written as [11]

$$X\,Z + c = 0 \qquad (1)$$

Where X= {x1, x2, x3 …, xn} is a weight vector and c a scalar (bias).

For 2-D it can be written as [11]

$x0 + x1\,z1 + x2\,z2 = 0$ where x0 = c is additional weight

The hyper plane defining the sides of the margin:

H1: $x0 + x1\,z1 + x2\,z2 \geq 1$ for yi = +1, and

H2: $x0 + x1\,z1 + x2\,z2 \leq -1$ for yi = –1

Any training tuples that fall on hyper planes H1 or H2 (i.e., the sides defining the margin) are support vectors [11].

If data were 3-D (i.e., with three attributes), then we have to find the best separating plane.

After we got a trained support vector machine, we use it to classify test (new) tuples. Based on Lagrangian[11] formulation, the MMH can be rewritten as the decision boundary.

$$d(ZT) = \sum i=1..L \; yi \propto i \; Zi \; ZT + c0 \qquad (2)$$

Where, yi is the class label of support vector Zi

ZT is a test tuple

∝i is Lagrangian multiplier
L is the number of support vectors

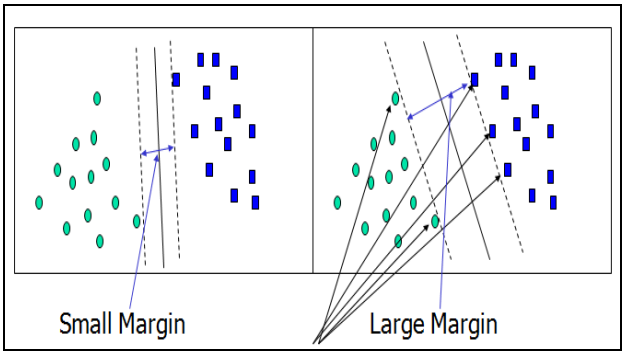


Figure 8: SVM margin and support vectors

EXPERIMENT and RESULTS

In this work the chain code method as discussed in section II was implemented in Matlab environment. The extracted chain code was normalized and used as features for two classifiers, namely, neural network and support vector machine.

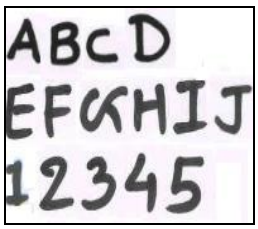
A. Implementation Results of ANN & Chain code based character recognition

In the implementation part, we have divided these 450 features in group of 15 with each group contains 30 features of letter A to T and digit 1 to 5 respectively. This matrix is used as a feature (figure 7) to train a neural network for training purpose.
At the other side for testing purpose, we have taken 30 different images. Binarization, segmentation is applied one by one on input image. Same feature matrix is prepared for all the segmented characters and applied these features to neural network for training and recognition. Here, we have used neural network classifier for the purpose of recognition. The overall accuracy of 80% was obtained for the test data using ANN.
The chain code method is used to recognize test input image shown in Fig.10 (a). The extracted chain codes for each character are shown in Fig. 10(b), while the final output stored in text file is shown in Fig.10(c).

```
   13   14   49   17   20   12   46   22
 7  16   35   12   10   18   28   17
 5  17   37   10   27   10   29   25
 4  23   30   10   31   12   25   26
 4  24   29   12   21   14   32   19
24   8   35    8   23    7   38    6
 4  15   30   10   20    9   26   20
 8  12   44   17   21    7   41   25
 8  12   44   17   21    7   41   25
 5  22   31   11   21   11   37   16
 5  27   25   15   22   14   34   19
          .    .    .    .    .    .    .    .    .    .    .    .
          .    .    .    .    .    .    .    .    .    .    .    .
            .    .    .    .    .    .    .    .    .    .    .    .

36  11   24   23   32    6   38   14
 50  11   18   23   42   12   24   16
 48  10   24   20   44    7   34   13
 36  11   24   23   32    6   38   14
```

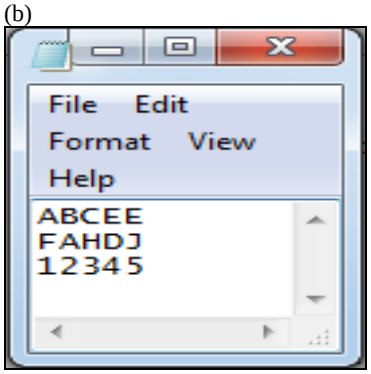
Figure 9: A subset of extracted chain code to be used for training neural network. Total 450 vectors each consisting of 8 features.



(a)

```
8    0   61   34    8    0   61   34
20    0    0   13   20    0    0   13
24    0   26   37   24    0   26   37
19    0    6   15   19    0    6   15
                .    .    .    .    .    .    .    .    .    .    .    .
                .    .    .    .    .    .    .    .    .    .    .    .
63    0   35   20   63    0   35   20
144   14   37   71  144   14   37   71
11    0   17   40   11    0   17   40
8    0   61   34    8    0   61   34
20    0    0   13   20    0    0   13
24    0   26   37   24    0   26   37
19    0    6   15   19    0    6   15
113   12   62   47  113   12   62   47
```

(b)



(c)

Figure 10 (a) input image (b) chain code of input image (c) output text file
It can be seen that out of 10 character and 5 digits in the input image 7 characters and all 5 digits were recognized properly.

B. Implementation Results of SVM & Chain code based character recognition

In the implementation part we have divided these 450 features in group of 15 with each group contains 30 features of letter A to T and digit 1 to 5 respectively as described in previous .This matrix is used as a feature to train a SVM for training purpose.
In the implementation part we have divided these 450 features in group of 15 with each group contains 30 features of letter A to T and digit 1 to 5 respectively as described in previous section. This matrix is used as a feature to train a SVM for training purpose.
At the other side for testing purpose we have taken 30 different images as described in previous section. Binarization, segmentation is applied one by one on input image. Same feature matrix is prepared for all the segmented characters and applied these features to SVM for training and recognition. We have used libsvm package [12] for the classification purpose. The

overall accuracy of 92% was obtained for the test data using SVM.

C Implementation Results of Correlation Coefficient based character recognition

We have used 10 samples of some characters and digits for testing. Table1 shows the different characters / digits and their correct recognition ratio found over testing samples. The recognition ratio is defined as,

Recognition ratio =

$$\frac{no\_of\_correct\_recognition}{total\_no\_of\_samples} * 100$$

D Comparison of Recognition using ANN and SVM classifiers
In table 1 we have listed different methods and accuracy. As shown in table we can easily say that overall accuracy of chain code (SVM) is good compare to chain code (NN) method. If we compare these methods on basis of

Table 1: Comparison of Overall Accuracy

Sr.no	Method	Structure	Accuracy
1	Chain code(Neural Network)	[8 30 6 25]	80%
2	Chain code(SVM)	-s 0 –t 2 –g 1 –c1	92%
3	Correlation Coefficient	-	80%

Table 2: Comparison of Individual Character Accuracy

Sr.no	Letter or Digit	Chain code-SVM Accuracy (%)	Chain code-NN Accuracy (%)	Correlation Coefficient
1	A	96	80	90
2	B	99	80	90
3	C	99	100	90
4	D	95	70	70
5	E	96	80	90
6	F	97	80	70
7	G	96	90	70
8	H	95	80	90
9	I	98	75	70
10	J	97	80	90
11	K	96	70	70
12	L	95	80	70
13	M	97	80	90
14	N	94	90	70
15	O	92	80	90
16	P	97	80	70
17	Q	95	90	70
18	R	93	90	90
19	S	97	80	70
20	T	94	80	70
21	1	97	80	100
22	2	96	90	100
23	3	95	80	100
24	4	99	80	100
25	5	96	80	75

training time then also SVM methods required less time compare to neural network. But drawback of SVM methods is we have to generate SVM format training and testing files, while in case of other methods it's not required. Now if we compare individual character accuracy then also chain code (SVM) gives good result compare to other method.

CONCLUSION

A simple and an efficient off-line handwritten character recognition system using a new type of feature extraction, namely, chain code is investigated. Selection of feature extraction method is most important factor for achieving high recognition ratio. In this work, we have implemented chain code based on 8-neighborhood feature extraction method. With the use of this obtained feature, we have trained the neural network as well as SVM to recognition character. In this work, in correlation coefficient method the template was considered to be the optical character, which is matched against the handwritten characters. A major drawback of the method is that the value of the correlation depends on the selected optical character template. If the style of writing handwritten character matches template then only a proper recognition ratio can be obtained. In the investigated work all two method showed the recognition of 80% or more.

REFERENCES

[1] S.. Mori, C.Y. Suen and K. Kamamoto, "Historical review of OCR research and development," Proc. of IEEE, vol. 80, pp. 1029-1058, July 1992.

[2] V.K. Govindan and A.P. Shivaprasad, "Character Recognition – A review," Pattern Recognition", vol. 23, no. 7, pp. 671- 683, 1990.

[3] H.Fujisawa, Y.Nakano and K.Kurino, "Segmentation methods for character recognition from segmentation to document structure analysis". Proceeding of the IEEE, vol.80, and pp.1079-1092. 1992.

[4] Yi-Kai Chen and Jhing-Fa Wang, "Segmentation of Single-or Multiple-Touching Handwritten Numeral String Using Background and Foreground Analysis", IEEE PAMI vol.22, 1304-1317, 2000.

[5] Pal, U. and B.B. Chaudhuri, "Indian script character recognition: A survey," Pattern Recognition", vol. 37, no.9, pp. 1887-1899, 2004.

[6] Ravi K Sheth, N.C.Chauhan, Mahesh M Goyani," A Handwritten Character Recognition Systems using Correlation Coefficient", selected in International conference V V P Rajkot, 8-9 April 2011.

[7] Dewi Nasien, Habibollah Haron, Siti Sophiayati Yuhaniz "The Heuristic Extraction Algorithms for Freeman Chain Code of Handwritten Character", International Journal of Experimental Algorithms, (IJEA), Volume (1): Issue (1)

[8] S. Arora" Features Combined in a MLP-based System to Recognize Handwritten Devnagari Character", Journal of Information Hiding and Multimedia Signal Processing, Volume 2, Number 1, January 2011

[9] H. Izakian, S. A. Monadjemi, B. Tork Ladani, and K. Zamanifar "Multi-Font Farsi/Arabic Isolated Character

Recognition Using Chain Codes", World Academy of Science, Engineering and Technology 43 2008

[10] R. J. Rummel, "Understanding Correlation" Honolulu: Department of
Political Science University of Hawaii, 1976.

[11] Jiawei Han and Micheline Kamber "Data Mining Concepts and Techniques", 2nd Edi, MK publication, 2006, pp 337-343

[12] Chih-Jen Lin,"A Library for Support Vector Machines", http://www.csie.ntu.edu.tw/~cjlin/libsvm/