International Journal of Electronics, Communication & Soft Computing Science and Engineering (IJECSCSE)
ISSN: 2277-9477, Volume 2, Issue 1

CR Methodology and HCR Recognition to Online Character Recognition.

M. V. Joshi. **J.B.Jadhav.** **V. R. Hire**

***Abstract*—In this paper on-line character recognition of single digit Deonagari numeral is done Here the character is written on pressure sensitive pen tablet. It has layers of conductive & resistive material with mechanical spacing between them. This digitizer sends the co-ordinates of pen tip to the host computer at regular intervals. Using the binary image of character various statistical features are extracted . Hu's moment invariants are used as features. The classification is done by Gaussian classifier. The handwritten data base is collected from different people by which the classifier is trained. The character is written on pen tablet, features of unknown character are extracted. These features are compare with features of data base images and output is observed on screen.**

***Keywords* - Pressure sensitive digitizer,CR methods, feature extraction, Gaussion classifier.**

I. INTRODUCTION

According to the mode of data acquisition, as online and off-line character recognition systems. In this paper the on-line recognition is used. The problem of recognizing handwriting, recorded with a digitizer, as a time sequence of pen coordinates is known as on-line character recognition. The digitizer uses pressure-sensitive tablets, which have layers of conductive and resistive material with a mechanical spacing between the layers.

There are also, other technologies including laser beam and optical sensing of a light pen. The on-line handwriting recognition problem has a number of distinguishing features, which must be exploited to get more accurate results than the off-line recognition. Handwritten Character Recognition (HCR) system typically involved two steps-- feature extraction in which the patterns are represented by a set of features and classification in which decision rules for separating pattern classes are defined.

Gaussian distribution classifier is used in present work with the help of mean and standard deviations of each feature the recognitions done.

The block diagram of classifier is shown below –

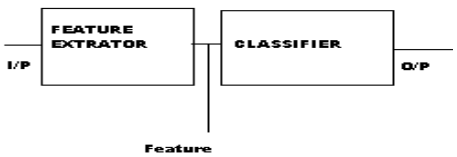


Figure 1. Classifier

AD Real valued feature vectors are ideal for statistical classifiers. In statistical pattern recognition a feature vector containing feature maps each pattern as a point in an N-dimensional feature space. Statistical information obtained from observations on a known set of representative patterns i.e. the training set is used to determine suitable features and boundaries between one class and another in order to maximize the recognition performance for each pattern class. If a given pattern is identical to a reference pattern, they will both be mapped to same points in feature space. Statistical classifiers need to be trained on a large data base of known patterns if they are to be effective. This large database is needed to define accurately the feature set and the classifier decision boundaries.

Type Style and Fonts

Wherever Times is specified, Times Roman or Times New Roman may be used. If neither is available on your word processor, please use the font closest in appearance to Times. Avoid using bit-mapped fonts if possible. True-Type 1 or Open Type fonts are preferred. Please embed symbol fonts, as well, for math, etc.

II. CR METHOLOGIES

A. On-lineRrecognition.

The progress in CR methodologies evolved in two categories according to the mode of data acquisition, as online and off-line character recognition systems. The digitizers are mostly electromagnetic-electrostatic tablets, which send the coordinates of the pen tip to the host computer at regular intervals. Some digitizers use pressure-sensitive tablets, which have layers of conductive and resistive material with a mechanical spacing between the layers. There are also, other technologies including laser beams and optical sensing of a light pen. The on-line handwriting recognition problem has a number of distinguishing features, which must be exploited to get more accurate results than the off-line recognition problem-

- Advantages of on-line -

1. It is a real time process**.** While the digitizer captures the data during the writing, the CR system with or without a lag make the recognition.
2. It is adaptive in real time. The writer gives immediate feedback to the recognizer for improving the recognition rate, as (s)he keeps drawing the symbols on the tablet and observes the results.
3. It captures the temporal and dynamic information of the pen trajectory. This information consists of the number and order of pen-strokes, the direction of the writing for each pen-stroke and the speed of the writing within each pen-stroke.
4. Very little pre-processing is required. The operations, such as smoothing, de-slanting, de-skewing, detection of line orientations, corners, loop and cusps are easier and faster with the pen trajectory data than on pixel images.
5. Segmentation is easy. Segmentation operations are facilitated by using temporal and pen-lift information, particularly, for hand-printed characters.

- Disadvantages of the on-line -

1. The writer requires special equipment, which is not as comfortable as pen and paper.
2. It cannot be applied to documents printed or written on papers.
3. Punching is much faster and easier than handwriting for small size alphabet such as English.
4. The available systems are slow and recognition rates are low for handwriting that is not neat.

- Applications of on-line –

This include small hand-held devices, which call for a pen-only computer interface and complex multimedia systems, which use multiple input modalities including scanned documents, speech, keyboard and electronic pen. On-line character recognition systems are useful in social environments where speech does not provide enough privacy. They provide an efficient alternative for the large alphabets where the keyboard is cumbersome. Pen based computers, educational software for teaching handwriting and signature verifiers are the examples of popular tools utilizing the on-line character recognition techniques.

B. Off-lineRecognition

Off-line character recognition is known as Optical Character Recognition (OCR), because the image of writing is converted into bit pattern by an optically digitizing device such as optical scanner or camera. The recognition is done on this bit pattern data for machine-printed or hand-written text. The research and development is well progressed for the recognition of the machine-printed documents. In recent years, the focus of attention is shifted towards the recognition of hand-written script.

- Advantage of the off-line --

These recognizers are to allow the previously written and printed texts to be processed and recognized.

- Disadvantages of the off-line --

1. Off-line conversion usually requires costly and imperfect pre-processing techniques prior to feature extraction and recognition stages.
2. The lack of temporal or dynamic information results in lower recognition rates compared to on-line recognition.

- Applications of off-line –

Some applications of the off-line recognition are large-scale data processing such as postal address reading, check sorting, office automation for text entry, automatic inspection and identification. Off-line character recognition is a very important tool for creation of the electronic libraries. It provides a great compression and efficiency by converting the document image from any image file format into more useful formats like HTML or various word processor formats. Recently, content based image or video database systems make use of off-line character recognition for indexing and retrieval, extracting the writings in complex images. Also, the wide spread use of web necessitates the utilization of off-line recognition systems for content based Internet access to paper documents.

III. HANDWRITEN CHARACTERS AND METHODS

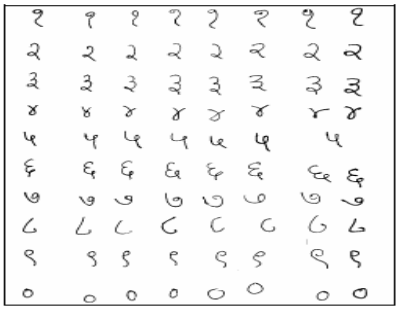
The script Devanagari originally belongs to Brahmi, which is considered purely Indian in nature. In the fourth century, from the northern branch of Brahmi, Gupta script was developed. Subsequently Kutil came out of Gupta script and Nagari script developed out o f Kntil in 8^{th} and 9^{th} century. The ancient Nagari script gave birth to modern Nagari, Gujrathi, Rajasthani, Marathi and Bangla scripts. Later on this modern Nagari script came to be known as “Devanagari”. Acharya Vinoba Bhave Calls it Lok-Nagari. According to him this script it used not only by one religion, caste or creed rather it has become the script of whole nation and of common people.

There are few opinions about its name “Devanagri”.

1. It was called Devanagari due to its exclusive use in Brahmins of Gujrath.
2. It was called Nagari for being prevailed in Nagars and Sanskrit was called voice of Devas, so Nagari was called “Devanagari”.
3. According to Shamshastri, idols of Devas were worshipped in symbols and these symbols were in the form of triangular yantras which were called Devanagar and since this script being developed form these symbols, so called as “Devangari”.
4. Another view-point is that it was prevalent in Devanagar area of Kashi hence named “Devanagari”.

It is indisputable that Devanagari has the most accurate scientific basis. For a long time; it has been the script of Indian Aryan languages. It is even now used by Sanskrit, Hindi, Marathi, Kokani and Nepali languages.

Presently, the Devanagari script is most scientific script found. Since every script is developed from Brahmi script, so Devanagari script has connection with almost every other script. Some Deonagari handwritten numerals are,



:

A set of seven invariant moments can be derived from the second and third moments.

$\phi_1 = \eta_{20} + \eta_{02}$

$\phi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta^2$

$\phi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2$ (11)

$\phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2$

$\phi_5 = (\eta_{30} + 3\eta_{12}) + (\eta_{30} + \eta_{12})\,(\eta_{30} + \eta_{12})^2$

$- 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03})\,(\eta_{21} + \eta_{03})$

$$[3(\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2]$$

$$\phi_6 = (\eta_{20} - \eta_{02}) + (\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2$$

$$+ 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}$$

$$\phi_7 = (3\eta_{21} - \eta_{02})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2$$

$$- 3(\eta_{21} + \eta_{03})^2] + (3\eta_{12} - \eta_{30})(\eta_{21} - \eta_{03})$$

$$[3(\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2]$$

This set of moments is invariant to translation, rotation, and scale change. As the property of invariant moment discussed above i.e. invariant under reflection, there is problem in recognition of 1, 9, 7 and 3, 6 because of their similarity under reflection. The recognition rate was found to be only 39% by using the seven invariants of each numeral. Therefore, the image is divided into 4 zones. Then we evaluated the invariant moments features of each parts. Thus in total there are 28 features. All these features are used in the recognition system. By including these 28 features the recognition rate can be enhanced

Another way of extracting features from image is dividing image into two zones. This method gives more information with different features.

IV. Image Compression

Principal Components Analysis (PCA). What is it? It is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. Since patterns in data can be hard to find in data of high dimension, where the luxury of graphical representation is not available, PCA is a powerful tool for analyzing data.

V. Recognition

Here the sample of hand written characters from 100 people are taken .Then moment features of every character are extracted .The mean and standard deviation of each feature is determined .thus a trained data base is prepared .

If unknown numeral feature is matched with trained data base numeral feature then it produces maximum membership value. Thus recognition is done. The formulae for mean and standers deviation are

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad s = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{(n-1)}}$$

VI. Result and Future Scope

In this work the data base is prepared near about with the help of 100 people with their hand written numerals. The data set contains varieties of writing styles .The result of recognition from different features is shown in table I

TABLE I
RESULT OF RECOGNITION

D. N…	Total I…	MIs t…	MIs+PCA t… S1 t… S2
०	100	61	78
९	100	64	92
२	100	82	88
३	100	88	93
४	100	95	90
५	100	81	88
६	100	95	91
७	100	74	68
८	100	67	97
९	100	94	89
		80.10	87.40

In this work the recognition of single numeral is done.

Also no method gives 100 percent result. So by improving the algorithm the 100 % recognition can be done.

References

[1] Nafiz Arica,fatos T.Yerman-vural,'An overview of character recognition focused on offline hand writing',IEEE trans.system,Man Vol.31,no.2(2001)

[2] Cho-Huak The, r.t.Chin, 'on image analysis by the methods of moments ', IEEE trans. Vol.10, No,4(1988)

[3] R.C.Gonzalez, R.E.Woods,S.L.Eddins,'Digital Image processing using matlab', Pearson Education,2004,pp 470-474

[4] R.J.Ramteke,P.D.Borkar,S.C.Mehritar 'recognition of Isolated marathi handwritten numerals',Proc.of.Int. Conf. on cognition and recognition,Mysore,india(Dec.2005),pp482-489.

[5] A tutorial on principal component Analysis, Lindsey I smith , Feb 26,2002.

Author's Profile

M. V. Joshi
Assistant Prof., SSVPS BSD
COE, Dhule
Ph No. 94222 89817
Email – mdhjoshi1@gmail.com

J.B.Jadhav
Assistant Prof., SSVPS BSD
COE, Dhule
Ph No. 9823080714
jayarp.2000@rediffmail

V. R. Hire
Assistant Prof., SSVPS BSD
COE, Dhule
Ph No. 9422214049
vrhsa@rediffmail