India Handwritten Digits Recognition

Iklaas A. Sultan
Computer Center
Mosul University

Received **Accepted**
12 / 03 / 2008 **15 / 10 / 2008**

الملخص:

يقدم البحث طريقة لتمييز الأرقام الهندية المكتوبة باليد. والتي تعتبر من أكثر الأرقام استخداما في العالم، التقنية المستخدمة في البحث هي تقسيم الصورة إلى قطاعات.

بعد إجراء بعض العمليات على الصورة (كالتقطيع، التحجيم و التنحيف) تقسم الصورة الى مجموعة من القطاعات كل قطاع يغطي زاوية ثابتة. مجموعة الخواص هي 24 وهي عبارة عن متجه المسافات والزوايا. لإجراء عملية التمييز استخدمت تقنية اقرب جوار لـK (K-Nearest-Neighbours). تم اختبار هذه الطريقة على 45 نموذج (رقم) خط يد لأشخاص مختلفين لكل رقم. حيث تم تقسيم عينة كل حرف إلى 20 نموذج تدريب و25 اختبار، النماذج المستخدمة في التدريب لا تظهر في الاختبار. يتراوح المعدل العام للتميز بنسبة 82.8% وهذا يعتبر أداء جيد جداً.

Abstract

An Optical Character Recognition (OCR) approach for handwritten Indian digit is presented in this paper, by using the proposed sector approach. In this approach, the normalized and thinned digit image is divided into sectors with each sector covering a fixed angle. The features totaling 24 include vector distances, angles. For recognition, the K-Nearest-Neighbours classifier is used. This method was tested using 45 patterns for each digit with different writers. The sample images were divided into 20 training and 25 test images. Images in the test set did not appear in the training sets. This method performs extremely well with recognition rates 82.8%. This is a very good performance.

1. Introduction

Optical Character Recognition (OCR) is a type of document image analysis where a scanned digital image that contains either machine printed or handwritten script is input into an OCR software engine and translating it into an editable machine readable digital text format (like American Standard Code for Information Interchange ASCII text)[1].

Text capture is a process to convert analogue text based resources into digitally recognizable text resources. These digital text resources can be represented in many ways such as searchable text in indexes to identify documents or page images, or as full text resources. An essential first stage in any text capture process from analogue to digital will be to create a scanned image of the page side. This will provide the base for all other processes. The next stage may then be to use a technology known as Optical Character Recognition to convert the text content into a machine readable format[1].

Document Image processing has been a frontline research area in the field of human machine interface for the last few decades. The need for efficient and robust algorithms and systems for recognition is being felt in Arabic language, especially in the postal department for sorting mail and for preserving out-of-print old books by digitizing them. Number recognition can also form a part in applications like intelligent scanning machines, text to speech converters, and automatic language-to-language translators[2].

The importance of recognizing Indian digits is also to be considered by Arabic non-speaking people such as Farisi, Curds, and Persians, and speaking who use the Indian digits in writing although the pronunciation is different[2].

For the above reasons and because of the benefits of (OCR) and after carefully studying the problems for recognized Indian digit, we construct recognition software[3].

This paper produce a technique for recognizing handwritten Indian digits using feature extraction properties which where analyzed by using statistical analysis correlation for the normalized and thinned image, the K-Nearest-Neighbours classifier is used for pattern classification methods.

2. Document Recognition Architecture:

Handwriting document recognition is the task of transforming image document into its symbolic representation; see Figure 1, [4].

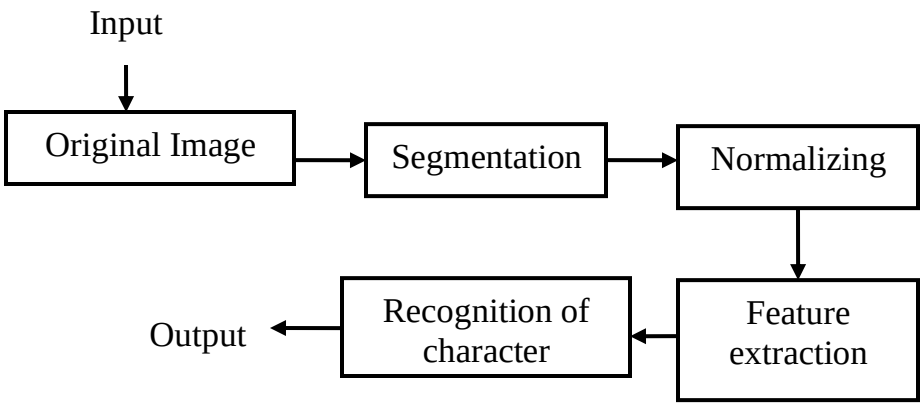

Figure (1): Handwritten recognition system

3. Preprocessing

It is necessary to perform several document analysis operations prior to recognizing digits in scanned documents. In any OCR system preprocessing includes the connection of segmentation and normalization. It is generally consists of a series of image-image transformation. It dose not increase our knowledge of the contents of the document, but may help to extract it[5].

3.1 Thresholding

The task of thresholding is to extract the foreground (ink, writing) from the background (paper)[6]. The histogram method is used for thresholding, the task of determining the threshold gray-scale value (above which the gray-scale value is assigned to white and below which it is assigned to black)[7].

3.2 Segmentation

To correctly recognize digits, we have to segment a binary image to set of images which only contain one digit. These digit images will be passed to the OCR module for recognizing. This is accomplished by examining the horizontal histogram profile. Line separation is usually followed by a procedure that separates the text line into words and characters. It focuses on identifying physical gaps using only the components [4].

3.3 Size Normalization

In OCR, very small and very large word or character images are often scaled to standard size, even though the outlines of characters of different

size in the same type face are not congruent. Size normalization is used to reduce the variation in size. Directly scaling all images to an identical size will result in significant deformation in many cases.

Size normalization for binary image *f(x,y)* applied in this OCR, so that the size of the rectangle circumscribing the pattern is 32 x 32 pixel. Consequently, the normalization image f'(x,y) is described as follow :

$$f'(x,y)=f(((\text{width}*x)/32)+\delta x\ ,\ ((\text{height}*y)/32)+\delta y) \qquad (1)$$

Where width and height are that of the pattern, respectively. Then δx and δy are the horizontal and vertical distance between the left-top corners of the image and the rectangle, respectively [5].

3.4 Thinning

Thinning is an image processing operation in which binary valued image regions are reduced to lines that approximate the centre lines, or skeletons, of the regions. The purpose of thinning is to reduce the image components to their essential information so that further analysis and recognition are facilitated. For instance, a line drawing can be handwritten with different pens giving different stroke thicknesses, but the information presented is the same [8].

Numerous algorithms have been proposed for thinning the plane region. This research uses an algorithm for thinning binary regions described in [9]. Region points are assumed to have 1 and background points to have 0. The method applied to the contour points of the given region (a contour point is any pixel of value 1 and having at least one 8-neighbor value 0). The thinning algorithm is as illustrated below:

Input: A digitized image I.
Output: A thinned image I.
Let N(p1) = p2+ p3 + p4 + p5 + p6 + p7 + p8
{ N(p1) number of non-zero neighbors of p1 }
Let S(p1) = { the number of 0-1 transitions in the ordered sequence of p3, p4, ., p9,p2 }.

(1) For every border pixel in the binary region, flag a contour point p for deletion if the following conditions are satisfied (b) S(p1) = 1;
(c) p2 . p4 . p6 = 0;
(d) p4 . p6 . p8 = 0;
(2) Delete the flagged points.

(3) For every border pixel in the binary region, flag a contour point p for deletion if the following conditions are satisfied:

(b) S (p1) = 1;

(c) p2 . p4 . p6 = 0;

(d) p4 . p6 . p8 = 0;

(4) Delete the flagged points, see Figure 2 [9].

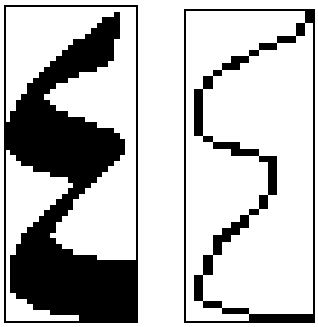

Figure (2): Thinning a handwritten for "4" digit.

4. Feature Extraction

The key issue of handwritten digits Recognition Software is feature extraction. Feature extraction abstracts high level information about individual patterns to facilitate recognition. Selection of feature extraction method is probably the single most important factor in achieving high recognition performance

4.1 Proposed Approach for Feature Extraction:

In this research structural features are explored in view of difficulty with topological and structural features for achieving high recognition rates.

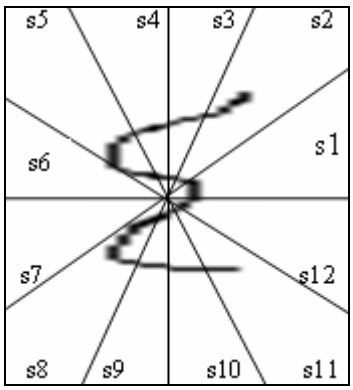

Figure (3) : Formation of sectors

In this approach, we consider the center of the digit matrix as the fixed point. This change makes the features more robust as they do not depend on the centroid. In this method, the normalized and thinned image is partitioned into a fixed number of sectors from the center of the image by selecting an angle. The number of sectors could be increased or decreased by changing the angle. However, from the experimentation, an angle of 30 degrees that leads to 12 sectors has been found to be an optimum choice. Here, we consider geometric features consisting of vector distances and angles. The pictorial representation of digit '4' subdivided into 12- sectors is shown in Fig.3. The first sector is from 0 to 30 degrees; the second sector is from 30 to 60 and so on. Once the digit is bifurcated into sectors, the portions lying in each sector is used for the extraction of features [11].

4.2 Extraction of Distance and Angle Features

Let *nk* be number of '1' pixels present in a sector *k*, with *k*=1,2,....,12. For each sector, the normalized vector distance, which is the sum of distances of all '1' pixels in a sector divided by the number of '1' pixels present in that sector is

$$D_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \left\{ (x_m - x_i)^2 + (y_n - y_i)^2 \right\}^{1/2} \tag{2}$$

Where, (*xi* , *yi*) are the co-ordinates of a pixel in a sector and (*xm* , *yn*) are the co-ordinates of the center of the digit image[10].

This normalized vector distance *Dk* is taken as one set of features. Next, for each sector the corresponding angles of pixels are also calculated. The normalized angle, *Ak* which is taken as another set of features, is calculated as:

$$A_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \tan^{-1} \left[\frac{y_n - y_i}{x_m - x_i} \right] \tag{3}$$

Both vector distances *Dk* and vector angles *Ak* constitute 24 features from 12 sectors.The relation for digit sample had been found through using vectors (*DK,AK*) properties which where analyzed by using statistical analysis correlation[11] . Table (1) represent the statistical data for each digit in the sample(1) see Figure (4).

Table (1) Statistical Data for Digit Sample in Figure (4)

Digit sample1	mean value **for Angle (0-30)**	mean value **for Angle (30-60)**	mean value **for Angle (60-90)**	mean value **for Angle (90-120)**	mean value **for Angle (120-150)**	mean value **for Angle (150-180)**	**Percent** excluded ca se percent	included case percent
0	15954.7684	15965.8675	15981.6572	15984.6503	15983.7493	15976.7612	4.5%	95.5%
1	17342.2392	17265.65528	17125.2342	17356.86735	17349.6331	17376.1637	5%	95%
2	16868.6348	16823.6081	16890.6320	16823.6574	16897.5462	1686.6193	2.2%	97.8%
3	16521.6574	16342.6574	16491.8754	16422.8936	16534.1245	16510.6234	3.1%	96.9%
4	18651.8957	18623.3546	18623.4561	18634.4643	18698.3452	18612.3234	4.2%	95.8%
5	17354.12345	17346.2345	17323.2453	17310.6756	17356.5675	17343.6765	4.7%	95.3%
6	16723.3444	16756.7632	16756.3392	16768.7643	16787.9812	16787.6452	3.3%	96.7%
7	17165.8732	17165.8190	17143.7684	171334.5431	17165.3219	17168.2387	2.8%	97.2%
8	16945.6531	16954.7684	16967.7612	16998.5627	16975.7684	16932.6512	1.1%	98.9%
9	17965.8794	17982.7563	17961.7863	17910.7562	17976.8385	17994.7684	2.8%	97.2%

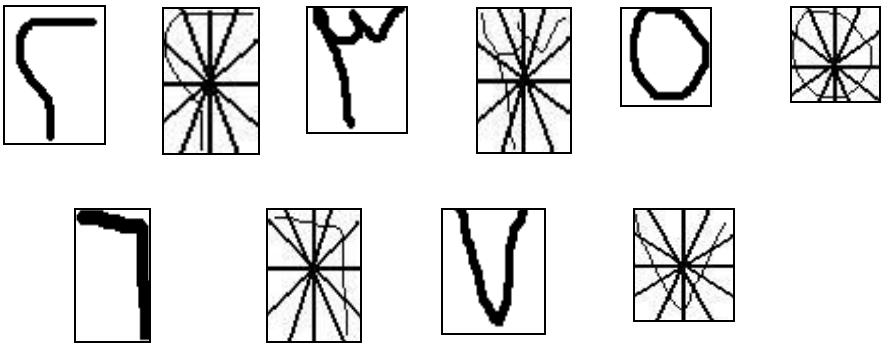

Figure (4): Sample 1 segmentation

5. The K Nearest-Neighbours Classifier

When using the *K* nearest-neighbours classifier (*K*NN), for each class *V* in the training set, the ideal feature vectors are given as *fv*. Then we detect and measure the features of the unknown digit (represented as *U*). To determine the class *R* of the digit we measure the similarity with each class by computing the distance between the feature vector *fv* and *U* the distance measure used here is the Euclidean distance [12]. Then the distance computed *dv* of the unknown digit from class *V* is given by

$$dv = \left[\sum_{j=1}^{N} \left(U_j - f_{vj} \right)^2 \right]^{1/2} \qquad (4)$$

where *J*=1,2,…, N (N is the number of the features considered).

The digit is then assigned to the class *R* such that:

$$d_R = \min(d_v) \quad (5)$$

where (*R*=1,…., no of classes)[12].

6. Experimental Result

A number of experimental were carried out to show the effectiveness of the proposed algorithms. Forty five samples for each digit were selected. For the purpose of the classification the sample images were divided into 20 training and 25 test images Figure(4). Images in the test set did not appear in the training sets. The images measure [175×175] with eight bits per pixel and the digits have been labeled manually. In this database, 45samples belong to category [0-9]. Table 2 lists the distribution of each category in our digits database. Some of samples from the database are shown in Fig. 4.

Table (2): Distribution of each category in test and recognition database

Category No.	Number of Digits tested	Number of digits recognized
0	**25**	**21**
1	**25**	**20**
2	**25**	**22**
3	**25**	**21**
4	**25**	**20**
5	**25**	**19**
6	**25**	**23**
7	**25**	**20**
8	**25**	**21**
9	**25**	**20**

This method performs extremely well with recognition rates 82.8% for 250 data base digit category.

7. Conclusion

In this paper we have presented a technique for recognizing handwritten Indian digit. A number of experiments have been conducted. The experiment used 25 samples for each digit. Features were extracted from digit image by divided it into sectors with each sector covering a fixed angle. The features totaling 24 include vector distances, angles. Recognition was performed through using K Nearest-Neighbours (K-NN) classification.

The results obtained were very promising and recognition as high as 82.8% was indicated.

All of this demonstrates that the new method is able to handle handwritten Indian digit task efficiently. It is a promising technique for recognition handwritten Indian digit.

References:

[1] Simon T., 2004, "Deciding whether Optical Character Recognition is feasible", Handprinted Arabic character recognition system". Pattern Recognition, Vo2. 10.

[2] Rama G., Ramakrishnan A. and Gupta D., 2002, "Parallel Processing in OCR – A Multithreaded approach", Proc. Tamil Internet, Singapore, P.165-170

[3] Kimura F, Shridhar M., 1991, "Handwritten Numeral Recognition Based on Multiple Algorithms", Pattern Recognition, V. 24 (Issue10), pp.969-983.

[4] Al-Zoubady, L. Alnsour, A. 2006, "Arabic Handwritten Characters Recognized by Neocognitron Artificial Neural Network" University of Sharjah Journal of Pure & Applied Sciences Volume 3, No. 2 p 1-17.

[5] AL-Zubaidy, L., 2002, "Arabic Machine Printed/Handwritten Character Recognition Using Neural Network", Mousl university.

[6] Amin A. Al-Sadoun, H. 1993, "Handprinted Arabic Character Recognition System". Pattern Recognition, Vol. 15.

[7] Tsai, Wen-Hsiang, 1984, "Moment-Preserving Thresholding: A New Approach" Document Image Analysis. p. 44-59.

[8] Kasturi, R. Govindaraje, V, 2002, "Document Image Analysis: A primer" Sadhana Vol. 27, Part 1, pp. 3–22.

[9] Amin, A. 1998, "Off-line Arabic Character Recognition the State of the Art" [review], Pattern Recognition, Vol. 31, pp. 517-530.

[10] Hanmandlu M., Mohan K. and Gupta V., 1997, "Fuzzy Logic Based Character Recognition", Proc. IEEE International Conference on Image Processing, pp. 26-29.

[11] M. Hanmandlu, M., Mohan and H. Kumar (1999), "Neural-based handwritten character recognition", Proceedings of International Conference on Document Analysis and Recognition, pp. 241-244.

[12] Said, H. E. S. Tan, T. N. and Baker, K. D., 1998, "Personal Identification Based on Handwriting", Pattern Recognition, vol.33, no.1, pp.149-160.