Hybrid Models for Offline Handwritten Character Recognition System Without Using any Prior Database Images

Kamal Hotwani, Sanjeev Agarwal and Roshan Paswan

Abstract In this paper a new method of classification is proposed by making hybrid models by using 3 different technique. One of them is correlation method, which use statistical template matching technique. Other one is principal component analysis in which for each image (character) there are some principal component, named Eigen value, and Eigen vectors. Third is Hough line detection technique, with the help of this we can find number of line segments in a character. Here with the experiments we can say that with the help of mixture of two or more different methods we can get better result. In this paper, we have implemented above techniques without using previous database of character images and getting 94.8% accuracy.

Keywords Offline characters recognition · Handwritten character recognition · Correlation · Hough line detection · Eigen value · Hybrid models

1 Introduction

Handwritten character recognition is one of the most emerging field of research. We can easily work on machine printed words/characters but handwritten characters are somewhat different. In handwritten, there are some challenges, like, uncertainty in writing styles, noise, two or more characters touching each others, some character are not connected, i.e., broken, improper scanning, lightning. It also depends on camera quality, angle at which image is taken, etc. It has number of applications

K. Hotwani (✉) · S. Agarwal
Malaviya National Institute of Technology, Jaipur, India
e-mail: kamalhotwani3@gmail.com

S. Agarwal
e-mail: sanagrawal@hotmail.com

R. Paswan
PEC University of Technology, Chandigarh, India
e-mail: roshanpaswan303@gmail.com

like recognizing formulas; perform analysis in banks, corporate, automatic form feeding, etc.

Character recognition is of two types, online character recognition, and offline character recognition. In online character recognition, there are two types of information, that is, spatial-luminance information. The luminance of character and the instance of time. Whereas in offline character recognition there is only one, factor that is, spatial information. Therefore offline characters recognition is little bit more complex than online character recognition. In this paper, we have taken some handwritten images of different persons that have upper case English alphabets and we try to recognize that characters via our technique. Handwritten recognition is now a day's one of the most challenging and emerging field. It can be used in various fields, number of areas. It can help to convert a handwritten data in image format to computer understandable characters. Here in this paper we have done experiments on uppercase English alphabets and converted in computer understandable format with 95% accuracy. There are some phases of our algorithm. They are as following:

1. Creating database by document itself
2. Preprocessing
3. Feature extraction or classification

These are explained in methodology section.

2 Previous Work

There are number of papers regarding character recognition but some of them are present in my following table, which acts as a torchbearer for us (Table 1).

Table 1 Literature survey

Author and year	Paper title	Contribution/work
Aravinda CV Dr H.N. Prakash @2014 IEEE	Template matching for Kannada handwritten recognition	In this paper author tried to recognize Kannad language via template matching (correlation) [1]
Nisha Sharma Bhupendra Kumar Vandita Singh @2014 IEEE	Recognition of off-line handwritten characters	In this paper characters, numerals and special symbols are also recognized via NN and SVM [2]
Mahesh Jangid Sumit Srivastava @2014 IEEE	Gradient Local auto correlation for handwritten devanagari characters recognition	They used further extension of HOG and SIFT and give 90–95% accuracy with GLAC-SVM [3]
Quan-Sn Sun Sheng-Gen Zeng De-Sen Xia @2004 IEEE	Feature Fusion Method based on canonical Correlation Analysis and HCR	In this paper they have used PCA, Eigen value and Eigen vectors to identify characters [4]

3 Challenges

There are some challenges in offline handwritten character recognition. Following are some glimpse of challenges to make concept clearer (Figs. 1, 2, 3, 4 and 5).

These are some of the problems faced during image processing. Here in this paper we have tried to solve these issues.

4 Methodology

See Fig. 6.

A. Creating database by document image itself
We have applied training and testing from the same image. We want to reduce the burden of extra database and training. We tried of minimize the difference between input characters and database characters. Therefore, for this we give handwritten A, B, C, D… characters in image format as input or database.
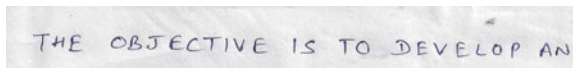
Fig. 1 Scanning quality
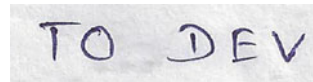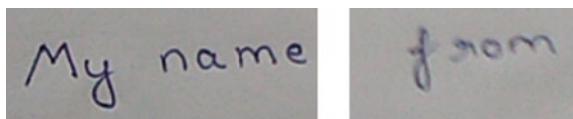

Fig. 2 Disconnected T and D
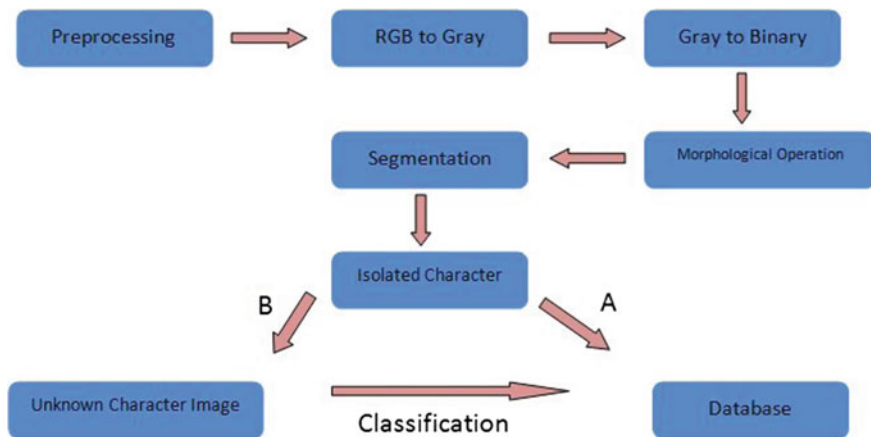

Fig. 3 Connected Y and S

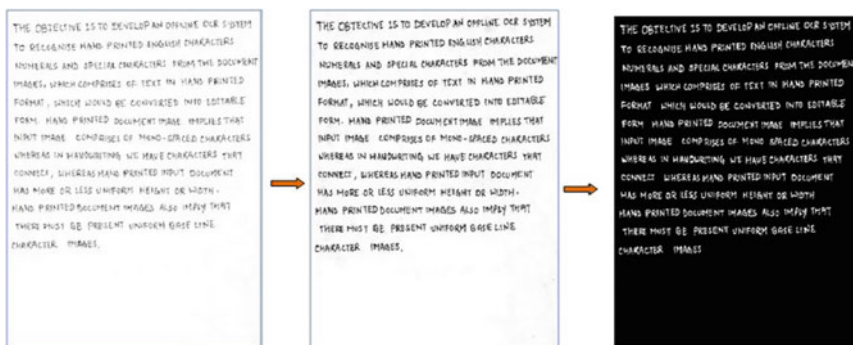**Fig. 4** Different types of intra-personal writing skills

Fig. 5 Brightness of image and blurred image

Fig. 6 Complete methodology **a** creating database by document itself **b** taking an unknown character from document image and classification is done by the database created

Fig. 7 Document image is enhanced and then binary conversion

The database characters are processed same as document image itself. The database image is processed, noise removal, colored image to grayscale image and then binary image. Then we try to identify different lines in an image and then different characters. All the individual characters displayed and we have to enter what is in the image. That is we are teaching computer that this is A, or B character. When we identify characters and make computer understand what is in the image. Than we make database of that character. This database of character is used to identify document characters.

B. Colored to grayscale conversion

We to retain the luminance while eliminating hue and saturation information. This method helps us to remove all unwanted information. After this, we also enhance the image for better output. We tried to remove noise and increase contrast (Fig. 7).

C. Binarization

In grayscale image, the value of each pixel is between 0 and 255. Zero represent black pixel and 255 represent white pixel. Now with the help of threshold value we will convert this grayscale image into binary image. The range of binary image is 0 to 1. Zero represent black pixel and one represent white pixel. We will also perform negative binarization, that is, 0 to 1 and 1 to 0.This helps us to visualize better and computer can work efficiently on this type of image.

D. Segmentation

We have to extract object of our interest from document image. This is required as classifier can only classify or identify isolated characters not a line or word as a whole [5]. First, we crop line segment from document image and after a single isolated character from that segmented line (Fig. 8).

E. Morphological operations

In morphological operations, we have used some edge detections techniques that is, canny edge detection to find edges and with the help of these edges, we can connect the non-touching part and fill it so that it can give better result as compared to non-fill characters (Fig. 9).

Fig. 8 Segmentation of line and then single isolated character

Fig. 9 Morphological operations, filling regions of character
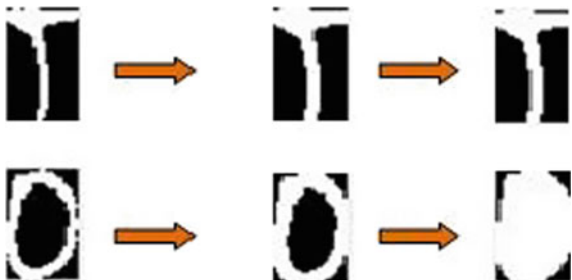
5 Classification

A. Correlation Coefficient

Correlation coefficient is a method to find relation between two matrixes A and B. In our case, we have two images with same size and in binary format and we are finding correlation coefficient that is how much image A is related with image B [6]. It is a quantative method that can find dependency between two 2-D array.

$$r = \frac{1}{n-1} \sum \frac{(x_i - \overline{X})(y_i - \overline{Y})}{s_x s_y} \tag{1}$$

where,

r   Relation between two matrix
n   Total number of pixels
x_i, y_i Value of ith pixel in x and y matrices

B. PCA: Principal component analysis

PCA is another statistical method converts observations into linearly uncorrelated data variable called as principal component. In this method we find principal components which are less than or equal to number of original variables [7]. Number of principal component less than or equal to Number of original variables.

In this, we have used Eigen value and Eigen vectors. The Eigen vector and Eigen value is computed of the document image character and then compared with all the character present in our database. This comparison is done by minimum square error method.

Characteristic Vector or Eigen Vector is a non-zero vector. It does not change its direction when linear transformation is applied.

$$T(v) = \lambda v \tag{2}$$

where,

T   Linear Transformation
T (v)   Scalar multiple of v
λ Eigen value (scalar value)

If T is a square matrix A, then

$$A v = \lambda v \tag{3}$$

Geometrically, Eigen vector corresponding to Eigen value, points in stretched direction by transformation and Eigen Value is quantative unit of stretching. We have calculated Eigen values of various isolated characters and database images and

then we find the minimum square error between them. The minimum error value is our expected character.

C. Hough Transform

Hough transform is a feature extraction technique. Here this method is used to detect lines in an image. First, we use canny edge detector and find out the edges in the image and after that, we find out the number of lines in an image. This line detection will give the number of line in an image and number of line is compared with the database images whose number of line we already know during database creation process.

The equation of straight line is as following:

$$r = x \cos \theta + y \sin \theta \tag{4}$$

where,

x, y Coordinates
r   distance from origin
θ   Angle between x-axis and line connecting origin and line segment.

6　Experiments and Analysis

For better results, we have mixed two or more methods to make hybrid model for classification.

A. Hybrid Model 1: Correlation + PCA

In this experiment, we have mixed two methods, Co-relation, and PCA. However, the problem is higher the co-relation better the relation and its value lies between −1 to 1. Moreover, for PCA we have used minimum square error technique that is minimum the error better the relation. Therefore, for making a hybrid method we have made a formula after doing some experiments and analysis [8], that is, co-relation has correct answer in second place or third place, if we arrange in decreasing order. Therefore, we have taken top five values and processed that value for PCA analysis and we get output with the following formula.

$$HC1 = r \times 100 \times K1 + (100 - pca\_cons) \times K2 \tag{5}$$

where,

HC1   Hybrid constant 1
r   correlation coefficient
pca_cons minimum square error of PCA
K1   preference constant for correlation method
K2   preference constant for PCA method

The value of r, pca_cons is calculated and put here. The value of K1 is 2.5 and K2 is 1.

B. Hybrid Model-2: Correlation + Hough transform

In this experiment, we have made a hybrid model that comprises of correlation and Hough transform. We tried to detect lines of each character and compared with characters present in the database. This can be helpful to detect or differentiate between M and N or T and I. Again, we taken top five values and processed that value for line detection. We get output with the following formula.

$$HC2 = r \times 100 \times K1 + K2 + \text{hough\_constant} \tag{6}$$

where,

HC2	Hybrid constant 2
r	correlation coefficient
K1	preference constant for correlation method
K2	preference constant for Hough transform method
hough_constant	if number of lines matches than some scalar value, else 0

The value of r is calculated and put here. The value of k2 is 1 here.

C. Hybrid Model-3: Correlation + PCA + Hough transform

In this experiment, we have mixed all three techniques and made a hybrid model that comprises of all three models.

We detected number lines of each character and compared with characters present in the database. Again, we taken top five values after correlation analysis and processed that value for PCA analysis and after that line detection. We get output with the following formula. The results obtained is better than other two models because numbers of lines and principal component have to find best match in top five expected character processed by correlation.

$$HC3 = r \times 100 \times K1 + (100 - \text{pca\_cons}) \times K2 + \text{hough\_constant} \tag{7}$$

where,

HC3	Hybrid constant 2
r	correlation coefficient
pca_cons	minimum square error of PCA
K1	preference constant for correlation method
K2	preference constant for pca method
hough_constant	if line matches than some scalar value, else 0

The value of r, pca_cons is calculated and put here. The value of K1 is 2.5 and K2 is 1.

Table 2 Results (accuracy) comparison

Persons/methods	Hybrid model-1 Corr + pca (%)	Hybrid model-2 Corr + Hough (%)	Hybrid model-3 Corr + pca + Hough (%)
Person x	93.5	93.3	94.5
Person y	96.23	93.9	95.8
Person z	93.1	92.67	94.14
Average	**93.1**	**93.29**	**94.82**

7 Results

In this paper, we have taken handwritten sample of some people and execute our hybrid models on different document images. I have taken three random images of three different persons, that is, three different handwriting and tabulated the results of experiment done (Table 2).

The hybrid model-3 gives us result up to 95%. Still there are some limitations and error but mainly error are due to handwriting error like two connected characters and disconnected single alphabet. The clear you write the better result you get.

8 Conclusion

In this paper, we tried to make a hybrid technology that can use multiple technique to get us a good result, that is, a better and reliable, handwritten character recognition technique. Hybrid method-1 gives result with 93.1% accuracy, hybrid method-2 gives 93.3% accuracy, and hybrid method-3 gives 94.8% accuracy.

Still there are some areas where future work can be done we can find out better algorithms or method which can give better results.

References

1. Aravinda, C.V., Prakash, H.N.: Template matching method for Kannada handwritten recognition based on correlation analysis. In 2014 International Conference on Contemporary Computing and Informatics (IC3I), pp. 857–861. IEEE (2014)
2. Sharma, N., Kumar, B., Singh, V.: Recognition of off-line hand printed english characters, numerals and special symbols. In: 2014 5th International Conference on Confluence The Next Generation Information Technology Summit (Confluence), pp. 640–645. IEEE (2014)
3. Jangid, M., Srivastava, S.: Gradient local auto-correlation for handwritten Devanagari character recognition. In: 2014 International Conference on High Performance Computing and Applications (ICHPCA), pp. 1–5. IEEE (2014)

4. Sun, Q.S., Zeng, S.G., Heng, P.A., Xia, D.S.: Feature fusion method based on canonical correlation analysis and handwritten character recognition. In: Control, Automation, Robotics and Vision Conference, 2004. ICARCV 2004 8th, vol. 2, pp. 1547–1552. IEEE (2004)
5. Dunn, C.E., Wang, P.S.P.: Character segmentation techniques for handwritten text-a survey. In: 11th IAPR International Conference on, Pattern Recognition, 1992. vol. II. Conference B: Pattern Recognition Methodology and Systems, Proceedings, pp. 577–580. IEEE (1992)
6. Prasad, J.R., Kulkarni, U.V., Prasad, R.S.: Offline handwritten character recognition of Gujrati script using pattern matching. In: 3rd International Conference on, Anti-counterfeiting, Security, and Identification in Communication, 2009. ASID 2009, pp. 611–615. IEEE (2009)
7. Deepu, V., Madhvanath, S., Ramakrishnan, A.G.: Principal component analysis for online handwritten character recognition. In: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004, vol. 2, pp. 327–330. IEEE (2004)
8. Kimura, F., Wakabayashi, T., Tsuruoka, S., Miyake, Y.: Improvement of handwritten Japanese character recognition using weighted direction code histogram. Pattern Recogn. **30**(8), 1329–1337 (1997)