Global Journal of Computer Science and Technology
Graphics & Vision
Volume 13 Issue 2 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

An Approach to Extract Features from Document Image for Character Recognition

By Mohammad Imrul Jubair & Prianka Banik
Ahsanullah University of Science and Technology, Bangladesh

Abstract - In this paper we present a technique to extract features from a document image which can be used in machine learning algorithms in order to recognize characters from document image. The proposed method takes the scanned image of the handwritten character from paper document as input and processes that input through several stages to extract effective features. The object in the converted binary image is segmented from the background and resized in a global resolution. Morphological thinning operation is applied on the resized object and then the technique scanned the object in order to search for features there. In this approach the feature values are estimated by calculating the frequency of existence of some predefined shapes in a character object. All of these frequencies are considered as estimated feature values which are then stored in a vector. Every element in that vector is considered as a single feature value or an attribute for the corresponding image. Now these feature vectors for individual character objects can be used to train a suitable machine learning algorithms in order to classify a test object. The k-nearest neighbor classifier is used for simulation in this paper to classify the handwritten character into the recognized classes of characters. The proposed technique takes less time to compute, has less complexity and increases the performance of classifiers in matching the handwritten characters with the machine readable form.

Keywords : *character recognition, morphological thinning operation, feature vectors, classifiers.*

GJCST-F Classification: *I.5.0*

Strictly as per the compliance and regulations of:

An Approach to Extract Features from Document Image for Character Recognition

Mohammad Imrul Jubair [α] & Prianka Banik [σ]

Abstract - In this paper we present a technique to extract features from a document image which can be used in machine learning algorithms in order to recognize characters from document image. The proposed method takes the scanned image of the handwritten character from paper document as input and processes that input through several stages to extract effective features. The object in the converted binary image is segmented from the background and resized in a global resolution. Morphological thinning operation is applied on the resized object and then the technique scanned the object in order to search for features there. In this approach the feature values are estimated by calculating the frequency of existence of some predefined shapes in a character object. All of these frequencies are considered as estimated feature values which are then stored in a vector. Every element in that vector is considered as a single feature value or an attribute for the corresponding image. Now these feature vectors for individual character objects can be used to train a suitable machine learning algorithms in order to classify a test object. The k-nearest neighbor classifier is used for simulation in this paper to classify the handwritten character into the recognized classes of characters. The proposed technique takes less time to compute, has less complexity and increases the performance of classifiers in matching the handwritten characters with the machine readable form.

Keywords : character recognition, morphological thinning operation, feature vectors, classifiers.

I. Introduction

Optical character recognition (OCR) has become an important field of research in the current growing period of technology. Automatic recognition of printed and handwritten information present on documents like cheques, envelopes, forms, and other manuscripts has a variety of practical and commercial applications in banks, post offices, libraries, and publishing houses. Basically OCR is a mechanism to convert machine printed or handwritten document file into editable text format [1]. The process of handwriting recognition involves extraction of some defined characteristics called features to classify an unknown handwritten character into one of the known classes. A typical handwriting recognition system consists of some steps, as like- preprocessing, segmentation, feature extraction, and classification [2]. Many methods have been proposed for recognizing the handwritten characters such as, HDCRGF [1], IHDCRFDHMM [2], HCRNN [3], EFHSNNHCR [4], and PABPNN [5] which can recognize the character in image by classifying them, but they take so much time and the methods are too complex and difficult to implement as well.

Recently SMHCR[6] has been proposed where a simplified technique is developed to recognize character from digital image. In that approach, the character object is segmented from the background and morphological thinning operation [7][8] is applied. After that the segmented image containing character object is partitioned into several cells. Feature value is estimated from each cell by calculating the proportion of the number of 0 and 255 intensity pixels. The estimated values for each cells are then stored in a vector and the vector is considered as a feature vector for that image.

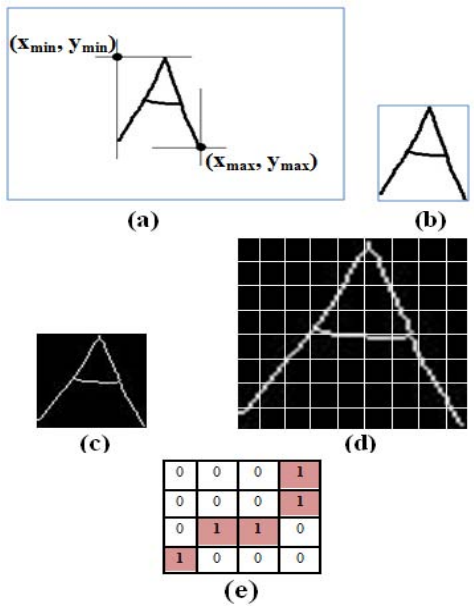


Figure 1 : (a) shows the extraction of character object from image and (b) is the extracted image, (c) is the result of morphological thinning operation on resized image, (d) shows the concept of partitioning the thinned image in same sized cells and (e) shows an example of calculating estimated value for a cell C where nw= 5 and nb= 11, so Pi = 5/11 = 0.454

K-nearest neighbor (KNN) [9] classifier is used here and the feature vectors are used to train the classifier. After training, the classifier is able to classify a

Author α : Lecturer, Department of Computer Science & Engineering, Ahsanullah University of Science and Technology, Bangladesh.
E-mail : mister_jubair@yahoo.com
Author σ : Lecturer, Department of Computer Science & Engineering, BGC Trust University, Bangladesh.

new test object into a recognized character. Fig. 1 shows the steps of SMHCR[6].

In SMHCR [6], the features are calculated from the proportion of 0 and 255 intensity pixels in a certain cell which is not efficient in all cases. Here the feature values are dependent on counting of pixels, rather than the shape of the object; though the shape is an important factor for recognizing a character. In this paper, a modified technique is proposed where shape of a character object is taken into account in order to estimate a feature value. The techniques searches for different shapes of joint in a character object and calculate the frequencies of their occurrence. The joint-shapes are pre-defined and their frequencies are considered as estimated feature values to be used in a suitable classifier.

II. Proposed Technique

Let **X** is the input character image with size m × n. Generally, documents are prepared by writing or typing on white paper. So, in this paper, we consider the background to be white and the foreground character objects to be black. **X** is converted into a binary image. So the background pixels will be the intensity of 1 and the foreground object pixels will be the intensity of 0. As mentioned earlier, the proposed technique seeks for predefined joint-shape occurrences.

Different kinds of joint-shapes are seen in character objects as indicated in Fig. 2 as an example.

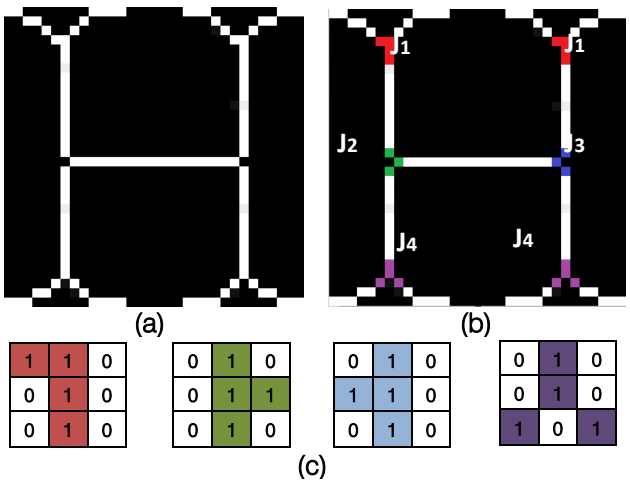


Figure 2 : (a) Shows a resized & thinned object (b) Indicates different joint-shapes J_i detected on the object which are numbered as J_1, J_2, J_3, J_4. and (c) Shows the template of joint-shapes J_1, J_2, J_3, J_4 (from left to right)

The concept of joint-shape detection can be illustrated with the example given in Fig.2. A character object is thinned and our technique searches for the four different predefined joint-shapes which are J_1, J_2, J_3, J_4. As we can see J_1 occurred for 2 times, J_2 occurred for 1 times, J_3 occurred for 1 times and J_4 occurred for 2 times. In this example, we will consider these frequencies to estimate feature values where-

$$\text{Frequency } (J_1) = F1 = 2$$
$$\text{Frequency } (J_2) = F2 = 1$$
$$\text{Frequency } (J_3) = F3 = 1$$
$$\text{Frequency } (J_4) = F4 = 2$$

Now let's consider a vector $F_X = [F1, F2, F3, F4]$
$= [2, 1, 1, 2]$

So, this F_X will be the feature vector for the image **X**. In practical case number of joint-shape template is more (J_1, J_2, J_3, J_4 ……. J_n) and in consequence feature vector contains more element such as **F = [**F1, F2, F3, F4,….,Fn**]** when number of joint-shape template, i = n. We can produce a histogram from the frequencies contained in a vector. Fig. 3 shows an example of histogram obtained by processing an image containing the character "A" where i = 24.

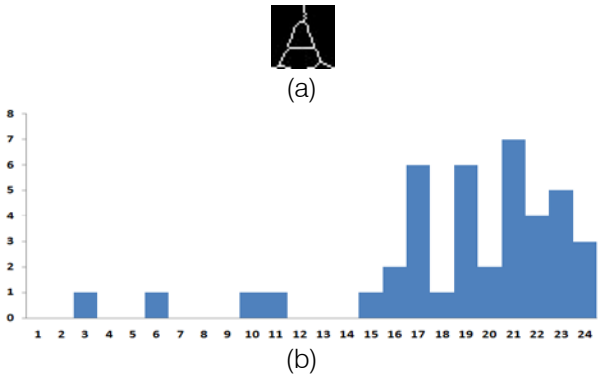


Figure 3 : (b) Shows the histogram of the feature vector obtained from (a) Image

In order to train a classifier, more images are processed to obtain feature vector and these vectors are passed into classifier to define individual classes. For an example, if we train classifier for the class "A" with 11 feature vectors from 11 objects, we get the following graph.

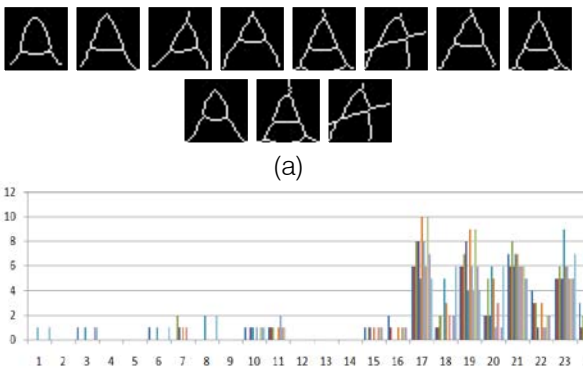


Figure 4 : (b) Shows Feature-vector histograms obtained from several training objects in (a)

III. Simulation

The proposed technique has been simulated using Matlab programming language. In our simulation, we have used 24 templates of joint-shapes. Fig. 5 shows the several templates.

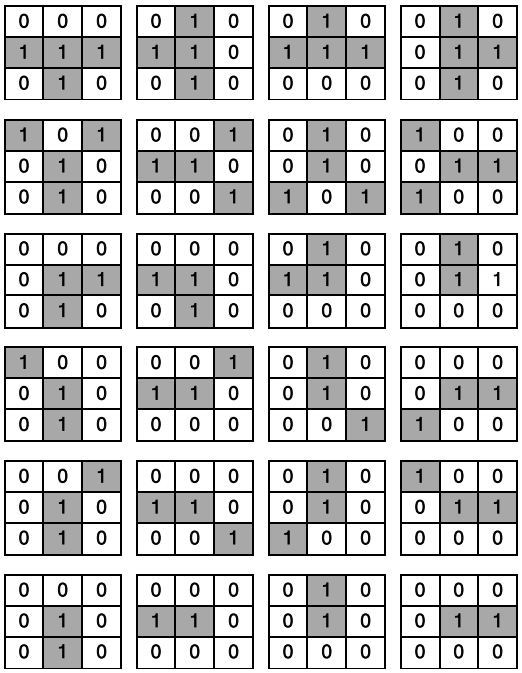


Figure 5 : Several joint-shape templates

The joint-shapes are used as window and they slide through the image to find out any match. In every matching the frequency values are incremented by 1.

Several character images are used to extract feature vectors and k-nn classifiers is used and is trained to determine the class of a testing object. Fig.6 shows some of the training images and feature vector histograms for several training classes are shown in Fig.7.

Figure 6 : Example of some training image objects

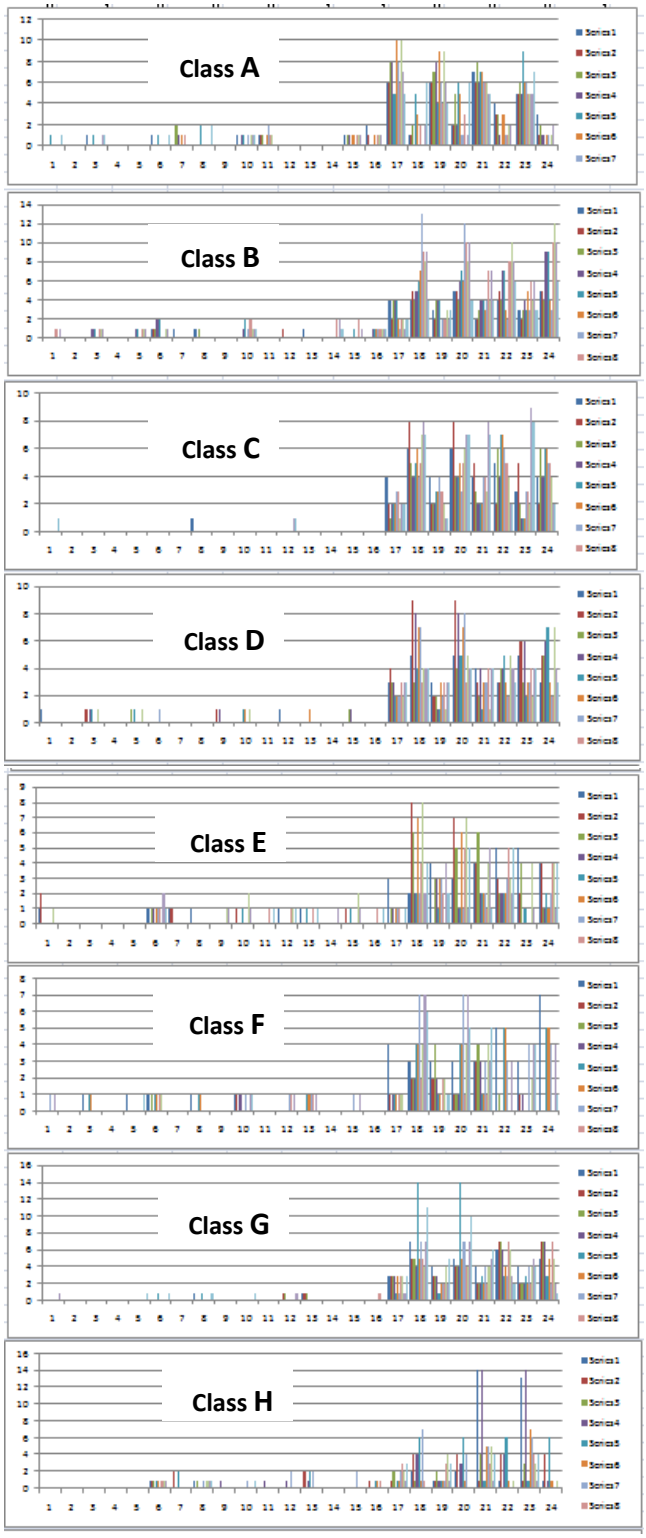


Figure 7 : Histograms of features vectors for several character classes (for simplicity only A to H are shown)

Total 650 number of sample images with different handwritten characters collected from different people has been tested using the proposed method and the result shows that the proposed method performs successfully to recognize handwritten

characters from document images and its average accuracy rate is 97.21%. The rate of success in recognizing sample images for different individual characters are shown in Fig 6 and a comparison between the proposed technique and SMHCR [6] is also presented.

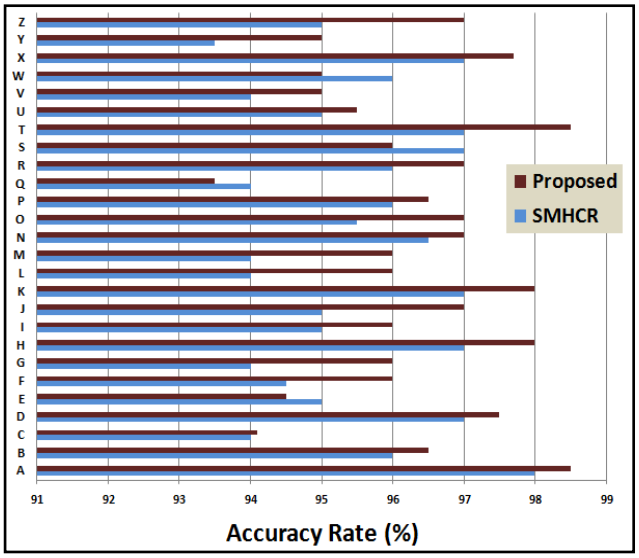
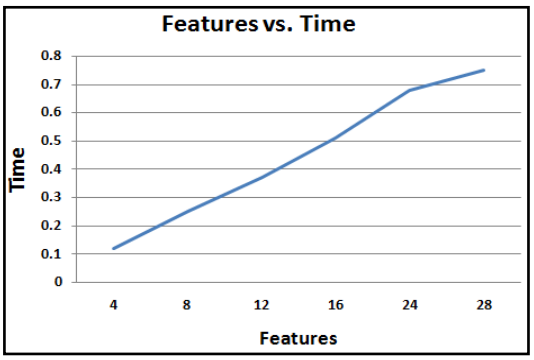
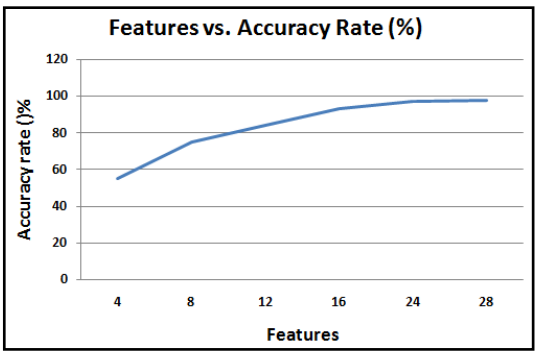


Figure 8 : Shows the accuracy rate of k-nn classifier in recognizing character using the proposed feature extraction method



(a)



(b)

Figure 9 : (a) Shows the number of feature element vs. running time graph and (b) Shows the number of feature element vs. accuracy rate graph

Fig. 9 shows some scenario related to feature vector's size. We can see that if we increase the number of elements in feature vector (i), the classifier needs more time to be trained (a) and as well as we will get more accuracy rate also (b).

IV. Future Plan

The proposed technique can be improved to make it more efficient. Predefined joint-shape templates can be selected carefully so that unused templates can be removed which will reduce time complexity. More powerful machine learning algorithms can be used in here in order to improve the recognition rate. Integration of this feature extraction method into the neural network is also a future plan of this work.

V. Conclusion

In this paper a method in presented to extract features from a document image. The features are extracted by seeking the occurrence of some joint-shapes in thinned object. The frequencies of occurrence are stored as feature elements in a vector. The feature vectors can be used through the classifiers in order to recognize a character object. The proposed feature extraction technique is less complex, easy to implement and integrate while recognizing the characters from document scanned image accurately.

References Références Referencias

1. Aggarwal, A., Rani, R. and Dhir, R. 2012. Handwritten Devanagari Character Recognition Using Gradient Features. International Journal of Advanced Research in Computer Science and Software Engineering, vol.2 - no.5, pp. 85-90.
2. Patil, S. B., Sinha, G.R. and Thakur, K. 2012. Isolated Handwritten Devnagri Character Recognition using Fourier Descriptor and HMM. International Journal of Pure and Applied Sciences and Technology, vol. 8 – no.1, pp. 69-74.
3. Patel, C. I., Patel, R. and Patel, P. 2011. Handwritten Character Recognition using Neural Network. International Journal of Scientific & Engineering Research, vol. 2 – no. 5, pp. 1-6.
4. Pawar, D. 2012. Extended Fuzzy Hyperline Segment Neural Network for Handwritten Character Recognition. Proceedings of the International Multi Conference of Engineers and Computer Scientists, vol. 1-no. IMECS 2012, pp. 43-46.
5. Kosbatwar, S. P. and Pathan, S. K. 2012. Pattern Association for character recognition by Back-Propagation algorithm using Neural Network approach, International Journal of Computer Science & Engineering Survey, vol. 3 – no. 1, pp. 127-134.
6. Jubair, M. I., and Banik, P. 2012. A Simplified Method for Handwritten Character Recognition from

Document Image. International Journal of Computer Applications, vol.51, no.14, pp.50-54.
7. Bansal, R., Sehgal. P. and Bedi. P. 2010. Effective Morphological Extraction of True Fingerprint Minutiae based on the Hit or Miss Transform. International Journal of Biometrics and Bioinformatics, vol. 4 – no.2, pp. 71-85.
8. Gonzalez, R.C., and Woods, R.E., 2004. Digital Image Processing (2nd edition), Pearson Education
9. Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z. H., Steinbach, M., Hand, D. J. and Steinberg, D. 2007. Top 10 algorithms in data mining. Knowledge and Information Systems, Springer-Verlag New York, Inc, vol. - 14, no. 1, pp. 1-37.
10. Charles, P. K., Harish, V. Swathi, M. and Deepthi, CH. 2012. A Review on the Various Techniques used for Optical Character Recognition. International Journal of Engineering Research and Applications (IJERA), vol. 2 – no. 1, pp. 659-662.
11. Baxes, G. A. 1994. Digital Image Processing: Principles and Apllications, John Wiley & Sons, New York.

This page is intentionally left blank