Automatic Identification of Handwritten Scripts

P. Ramanathan

Department of EEE, Bharath University, India

Abstract: Automatic identification of handwritten script facilitates many important applications such as automatic transcription of multilingual documents and search for documents on the Web containing a particular script. The increase in usage of handheld devices which accept handwritten input has created a growing demand for algorithms that can efficiently analyze and retrieve handwritten data. This paper proposes a method to classify words and lines in a handwritten document and deals with filtration of noisy data using mean and median filters. The structural features are extracted by using the special algorithm. These features are classified based on different spatial and temporal features extracted from the strokes of the characters. This method will be worked efficiently compared to the conventional character recognition soft wares.

Key words: Automatic transcription • Paper proposes • Structural features

INTRODUCTION

Handwriting recognition has always been a challenging task in pattern recognition. Many systems and classification algorithms have been proposed in recent years [1]. Techniques ranging from statistical methods such as PCA and Fisher discriminate analysis to machine learning like neural networks or support vector machines (SVMs) have been applied to solve this problem. But since handwriting depends much on the writer and because we do not always write the same character exactly in the same way, building a general recognition system that would recognize any character with good reliability in every application is not possible [2]. Typically, the recognition systems are tailored to specific applications to achieve better performance. In particular, unconstrained handwritten digit recognition has been applied to recognize amounts written on checks for banks or zip codes on envelopes for postal services (the USPS database). In these two cases, good results were obtained [3-6].

The performance of a classifier can rely as much on the quality of the features as on the classifier itself. A good set of features should represent characteristics that are particular for one class and be as invariant as possible to changes within this class [7-10]. Commonly used features in character recognition are: zoning feature, structural feature, directional features, crossing points and contours. A feature set made to feed a classifier can be a mixture of such features. Besides, to reduce the size of the feature set, feature subset selection can be applied to the extracted features. In handwriting recognition, features are created from knowledge of the data. But in some other applications, one may not have this knowledge that can be used to develop feature extractors. Another approach to this problem is to consider the feature extractor as a black box model trained to give relevant features as outputs with no prior knowledge on the data.

The earlier systems grouped under optical character recognition (OCR) could recognize only the printed or handwritten numerals of 2xed size and fonts. But, 100 percent recognition rate is beyond the reach of these systems as reported in the literature. Therefore, the present study aims at producing an accurate system targeting 100 percent recognition in face of varied size, shapes and fonts. An unconstrained and written digit recognition system can be divided into several stages: preprocessing (filtering, segmentation, normalization and thinning.), feature extraction (and selection), classification and verification. This paper focuses on feature extraction and classification. Since many classifiers cannot efficiently process the raw images or data, feature extraction is a preprocessing step that aims at reducing the dimension of the data while extracting relevant information.
people have special feelings about things. Losing things connected with special feelings causes the most hurt and the most anger. And each of us has many things in our lives that we invest with special feelings. A baseball player may have a special bat, he thinks it is lucky but he thinks that without that bat, he would not get to hit the ball as far. Part of his picture of himself as a ballplayer is tied up with that bat. Can you imagine how angry he would be if someone accidentally broke it? It would be like breaking his picture of himself. Imagine that a girl has a dog she loves it is just an ordinary dog nothing valuable about it. But would the girl sell it, no honoree you can't put a value on feelings.

**Fig. 1: Scanned image**

**Preprocesing:** Vijayaraghavan et al. [11]. In the preprocessing section, we have to collect the data which has to be recognized by scanning the original paper. Every scanned paper will have unwanted data.

This is shown in above Fig. 1. So we have to remove the unwanted data and we should improve the quality of scanned image by using image enhancement techniques. To enhance the image we are following the below two methods.

- **Histogram Processing**
- **Neighborhood Processing:**

  The main concept of neighborhood processing is the filtering. By using the median filter we can improve the image quality better compared to mean filter.

  After filtering the unwanted parts of the scanned image were removed by using a special algorithm. In this manner we can retrieve the better image.

**Feature Extraction:** In this section, each sample or pattern that we attempt to classify is either a word or a set of contiguous words in a line. Vijayaraghavan et al. [12]. Totally 32 features are extracted to gain accuracy. The extracted features are the classified by special algorithm adopted in this paper. The algorithm shows 98 percent accuracy in its classification.

**Line and Word Detection:** The data available to a script recognizer is usually a complete handwritten page or a subset of it. To recognize the script of individual lines or words in the page, we first need to segment the page into lines and words. The problem of text line identification in online documents has been attempted before. To identify the.

Individual lines, first the interline distance is estimated. The interline distance, d, is defined as the distance between successive peaks in the autocorrelation of the y-axis projection of the text. Lines are identified by finding valleys in the projection, keeping the interline distance as a guiding factor. To avoid local minima, we choose only those points, which have the smallest magnitude within a window of width d, as valleys. Once the line boundaries are obtained, the text is divided into lines by collecting all the strokes which fall in between two successive line boundaries. The temporal information from stroke order is used to disambiguate strokes which fall across line boundaries and to correctly group small strokes, which may fall into an adjacent line. Temporal information is also used to split lines in pages with multicolumn text (e.g., the document in Fig. 1). Fig. 2a shows the output of our line detection algorithm for part of the multicolumn document in Fig. 1. A word is defined as a set of strokes that overlap horizontally. The segmentation of a line into words is done using an x-axis projection of the text in the line. The valleys in the projection are noted as word boundaries and the strokes which fall between two boundaries are collected and labeled as a word. The minimum width of a valley in the projection for word segmentation was experimentally determined as 30 pixels for the resolution of the digitizing device used (0:1 mm). Fig. 2b shows the output of our word detection algorithm for the document in Fig. 1.

**Classification of Contiguous Words and Text Lines:** In many practical applications, contiguous words belonging to the same script are available for classification. Vijayaraghavan et al. [13]. We expect that the script recognition accuracy will improve as the number of consecutive words of text in a test sample increases. In the case of handwritten script recognition, this boils down to the number of words of text that is required to make an accurate prediction. The plot in Fig. 3 shows the increase in accuracy of the combined classifier as a function of the number of words in a test sample.

A set of words was considered as a single pattern for classification in this case. We notice that with five words, we can make a highly accurate (95 percent) classification of the script of the text. The script classification accuracy improves to 95.5 percent when we use an entire text line, consisting of an average of seven words. The error in accuracy estimate is about 1 percent...
as indicated by a standard deviation of 0.5 percent. The accuracy of prediction of the script of a single word also depends on the length of the word. A measure of word length, which can be employed in the case of online data, is the number of strokes in the word.

After classifying the text we can store all the characters in a word format file in different location this is done in the special algorithm. The classification improves considerably as the number of strokes in the word increases (up to 89 percent for 5-stroke words). These results give us an indication of the improvement in performance as the amount of data increases.

**CONCLUSION**

We have presented a script identification algorithm to recognize handwritten script in an online document. The aim is to facilitate text recognition and to allow script-based retrieval of online handwritten documents. The classification is done at the word level, which allows us to detect individual words of a particular script present within the text of another script. The classification accuracies reported here are much higher than those reported in the case of script identification of offline handwritten documents, although the reader should bear in mind that the complexities of the two problems are different.

Vijayaraghavan et al. [14], One of the main areas of improvement in the above algorithm is to develop a method for accurately identifying text lines and words in a document. We are currently working on developing statistical methods for robust segmentation of online documents. The script classification algorithm can also be extended to do page segmentation, when different regions of the handwritten text are in different scripts [15-17].
REFERENCES