A System for Handwritten and Printed Text Classification

Bala Mallikarjunarao Garlapati  
*TCS Innovation Labs*  
TATA Consultancy Services  
Hyderabad, India  
balamallikarjuna(dot)g(at)tcs(dot)com

Srinivasa Rao Chalamala  
*TCS Innovation Labs*  
TATA Consultancy Services  
Hyderabad, India  
chalamala(dot)srao(at)tcs(dot)com

Abstract—An optical character recognition (OCR) system recognizes either printed or handwritten text. Hence it is required to separate machine printed text from handwritten text in scanned documents before feeding it to an OCR system. We can discriminate these two types of text word images by their visual impression and shape structures. The intensity values distribution features gives us the visual impression and the shapes can be represented by the structural features. This paper proposes an approach for machine print and handwritten text classification at word level using intensity and shape structural features of scanned text. The proposed method achieved impressive classification efficiency on IAM dataset.

Keywords—Printed text, Handwritten text identification, Intensity features, Structural features, Support Vector Machine (SVM)

I. INTRODUCTION

Processing electronic or scanned documents like bank checks, financial instruments etc. require fine grained access control over certain fields for fulfilling security and privacy concerns. Some of these tasks require automatic localization and identification of fields. Identification of these fields require OCR to recognize the list. But these scanned documents may contain both printed and handwritten text.

The general block diagram of OCR system shown in Figure 1. An Optical character recognition (OCR) system scans the printed or hand written text documents and converts it into machine encoded text, then speak it or store it in a computer. In the machine encoded text conversion process, a separate OCR technology for machine print and handwritten text is required to make an efficient and robust system.

Separation of mixed text is required for using respective efficient OCR system. Visual appearance and structural properties of machine printed text and handwritten text are different. In this paper we proposed to use the best of these features and hence these are used for classification. For classification of printed text and handwritten text, support vector machine is used.

![Figure 1. Flow diagram of a generic OCR system](image)

The paper is organized in the following manner: Related work on printed and handwritten classification methods is discussed in section II. The visual and structural differences between printed and handwritten text and proposed new set of features used for classification are explained in section III. In section IV includes the experiments conducted and the corresponding results for proving the classification efficiency. Summarized the conclusion in section V.

II. RELATED WORK

Handwritten and Machine print text can be seperated using various features such as shape orientation. As shape of the printed text is unfluctuating compared to handwritten text shape can be used one of the features, Radon transform can be used to represent shape structural properties. A set of features based on Radon transform are generated for printed and handwritten text classification in [1]. In [2], A gray level 6-elements feature set is estimated and these features are ranked based on discriminatory capabilities. A decision tree classifier has been designed using feature set value boundaries for machine print and handwritten text classification. But the margin between the boundary values is small and the boundaries are crossing (overlap) each other for most of the features.

The Chain code method which is explained in [4] deals with the shape structural properties of given input text. The printed text has many straight strokes and handwritten text contains inclined strokes in different directions. The classification accuracy is low for scanned document images. In [6], data mining techniques on the classification step. A new set of features are proposed and mining classification rules are used to discern printed text from handwritten. In [7] a two step process method is proposed which has patch level separation and pixel level separation. It identifies three different categories of classes machine printed text, handwritten and overlapped text using G-means based classification. Further the overlapped text is separated using Markov Random Field based classification.

An overall system to localize the text areas and to divide them into text lines explained in [8]. They used structural characteristics to classify handwritten and printed text lines. To discriminate between machine printed and handwritten text...
Arabic and Latin scripts, different sets of features were used in [5]. Structural features which are intrinsic to Arabic and Latin scripts are also used here. Zheng and Li explained in [9] about handwriting and printed text segmenting and identification in noisy document images. The novel aspect in their work is considering the noise as a separate class and modeled the noise using selected features. The straightness of vertical and horizontal oriented lines and the symmetry with respect to points in the character image, are used in [11] for machine print and handwritten character discrimination.

There have been some works aiming for Indian scripts. Indian script identification using features based on topological, component and fractal is explained in [12]. Handwritten and machine print text lines identification of two Indian scripts Bangla and Devnagari explained in [10]. Vertical and horizontal projection profiles and their runs used as feature to discriminate between them. Statistical texture features like mean, standard deviation and entropy are used for printed and handwritten text classification in Indian Script languages in [3].

III. PROPOSED METHOD

The difference between handwritten and machine print text can be easily recognized by a human brain. But recognizing the difference by a machine requires some intelligence to understand features unique to each type of text. Some of the visual and structural differences listed in table I.

In this paper we propose a method using several features and their combinations which represent unique characteristics of machine printed and handwritten text, so that either can be easily identified in mixed text document images. The overview of system block diagram showed in Figure 2. It has three steps to process the classification. Text localization, feature extraction and classification. The scanned documents are first segmented into word images (both machine print or handwritten). Then different types features are extracted from the collected word images. These features are explained in the below section. Multiple features are concatenated and are learnt using a SVM classifier to differentiate between the handwritten and machine printed texts.

Figure 2. An overview of Machine print and Handwritten text classification system

A. Text localization

We use handwritten and machine printed text document images are collected from IAM dataset [13]. These documents are segmented into words using the methods mentioned in [3]. It has three steps process. 1) using Otsus

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Machine printed text</th>
<th>Handwritten text</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Texture (feel and appearance) is different from the Hand-written.</td>
<td>Texture (feel and appearance) is different from the Machine printed.</td>
</tr>
<tr>
<td>2</td>
<td>Pixel intensity values are almost similar for a whole word, except scanning noise like ‘dots’.</td>
<td>Pixel intensity values are not that much similar for a whole word.</td>
</tr>
<tr>
<td>3</td>
<td>More straightness in each character, compare to hand-written text.</td>
<td>Not much straightness in each character, compare to machine-printed text.</td>
</tr>
<tr>
<td>4</td>
<td>Character shape is unique in each font type.</td>
<td>The text shape depends on each individual person.</td>
</tr>
<tr>
<td>5</td>
<td>The size of characters (height and width) is constant.</td>
<td>The size of characters (height and width) is not constant.</td>
</tr>
<tr>
<td>6</td>
<td>The gaps between characters in a word are almost constant.</td>
<td>The gaps between characters in a word are not constant (irregular).</td>
</tr>
<tr>
<td>7</td>
<td>Number of black pixels to represent a character is constant for a particular font type and font size.</td>
<td>Number of black pixels to represent a character is not constant.</td>
</tr>
<tr>
<td>8</td>
<td>Line straightness is more here.</td>
<td>Line straightness is less here.</td>
</tr>
<tr>
<td>9</td>
<td>The variance in intensity values are less compare to handwritten.</td>
<td>The variance in intensity values are more compare to machine-printed.</td>
</tr>
<tr>
<td>10</td>
<td>The stroke width is constant for a particular font type and font size.</td>
<td>The stroke width is not constant.</td>
</tr>
<tr>
<td>11</td>
<td>Density of the pixels is more compare to handwritten.</td>
<td>Density of the pixels is less compare to machine-printed.</td>
</tr>
<tr>
<td>12</td>
<td>Looks more smoother compare to handwritten. (particularly at curved shapes)</td>
<td>Looks not smoother compare to Machine-printed.</td>
</tr>
<tr>
<td>13</td>
<td>Variations in shapes (circles, eclipse...) are less.</td>
<td>Variations in shapes (circles, eclipse...) are more.</td>
</tr>
</tbody>
</table>

threshold, binarize the document. 2) apply morphological operations to remove unwanted components in the document like quotation marks, dots, colons... . 3) apply connected component rules to do word separation in the scanned document.

B. Feature extraction

The main difference between machine printed and handwritten text is their visual appearance and shape structures. These properties are exploited for feature extraction. Visual impression can be represented by their gray level intensity values distribution. The intensity parameters like mean, standard deviation, otsu’s threshold, local maxima and upper quarter intensity are used to show the visual impression differences between the two types of text. These features are previously used in [2]. Shape structural differences can be represented by uniformity in pixel intensities and physical sizes. Density and variance of intensity values represent the structural differences between the two types of text. The variations in all these features for both printed and handwritten texts showed as graphs in next section IV.
All the feature values are combined into a single array to form a final feature vector of a word image. The feature vectors for all the word images are computed and stored for final classification. Each word image and hence its feature vector is assigned a label based on their category (1: Machine print text, -1: Handwritten text). Then we train SVM classifier with the collected feature vectors and labels.

The trained SVM classifier is used to classify test word images of machine print and handwritten text.

IV. EXPERIMENTAL RESULTS

In this section, experimental results of the proposed method are discussed. IAM dataset text document images contain both handwritten and printed text. These text documents are segmented into word images. These segmented word images are used in our experiments. Various features described in previous section are extracted from these word images and these are used to analyze the differences between machine print and handwritten text. The following figures show how the proposed features can distinguish between machine print and handwritten text.

Pixel density represents how much area of black binary pixels occupied in the total word image area.

\[
\text{Pixel Density} = \frac{\text{Total No.of the Active Pixels}}{\text{Area of the Word Image in Pixels}}
\]  

The pixel density variations in both types of text showed in Figure 3. From this we can see that most of the values of machine printed are in between 0.2 and 0.45. There is no clear separation between machine printed and handwritten text density values, but this feature can be used to augment the classification efficiency.

Variance of pixel intensity values in word images corresponding to both types of text showed in Figure 4. Handwritten values exhibit higher variances compared to machine print text. Hence second moment of intensity values is considered as a feature for separating handwritten and machine print text.

Word mean represents the average gray value of a given word image. The word mean feature values in both types of text in Figure 5. Mean of the machine print text is well above 200 for most of the word images and the handwritten text has its mean values below it. Hence this word mean can be useful in distinguishing between handwritten and machine print text.

Standard deviation of pixel intensities represents how the gray level values spread out from the mean. The standard deviation for both types of text showed in Figure 6. As we can see from the figure, there is a clear separation between machine print and handwritten text standard deviation feature values. Most of the machine print values are more than 75 and handwritten written values are varying from 30 to 75. We noted that most of the classification efficiency is coming from this feature.

Figure 3. Density feature comparison of Machine print and Handwritten text word images

Figure 4. Variance feature comparison of Machine print and Handwritten text word images

Figure 5. Mean feature comparison of Machine print and Handwritten text word images

Figure 6. Standard deviation feature comparison of Machine print and Handwritten text word images
Upper quarter region intensity represents how the pixel intensities occupied in the upper quarter portion of the word image. From the machine print text images it is observed that the number of pixels with active intensity values in the upper quarter region are more while for handwritten text there are very few pixels that has active intensity values. The plot of the number of upper pixel intensity values proves our hypothesis. The Upper quarter region intensity feature in both types of text showed in Figure 7. Most of the values of machine print text are concentrated in between 70-90 and handwritten values are in between 50-70 . This feature is useful for classification of two types of text.

Figure 7. Upper quarter region intensity comparison of Machine print and Handwritten text word images

Otsu’s threshold in general is used for separating background with foreground. The threshold value that separates dark and light regions of word image. As the Otsu’s threshold would be different for handwritten and machine print text images due to their intensity variations. Hence this can be used as a feature for machine print and handwritten text classification.

The otsu’s threshold values in both types of text plotted in Figure 8. From the figure we can observe a clear separation between machine print and handwritten text otsu’s threshold values. Almost all of the Otsu’s threshold for machine print text are less than 150 and handwritten text has Otsu’s thresholds above 150. Hence this feature can discriminate machine print and handwritten text words efficiently.

Figure 8. Otsu’s threshold feature comparison of Machine print and Handwritten text word images

Number of local maximas represents the straightness variations in a given word image. The local maxima feature values in both types of text showed in Figure 9. Most of the values of machine print text are concentrated in between 60-80 and handwritten values are in below 60. This feature is useful for classification of two types of text.

Figure 9. Local maximas comparison of Machine print and Handwritten text word images

Finally all these features are concatenated to arrive at a feature vector that represents each word. We tested the proposed features on IAM dataset. IAM dataset [13] contains handwritten text scanned images which can be used for training and testing of handwritten text detection problems. 657 different writers produced 1539 pages of hand written text forms and these forms are scanned at 300 dpi resolution and saved as PNG format files. The database consists a total of 115320 labeled words and these words have been extracted from the scanned text images using a word segmentation scheme.

We use 10 fold cross validation technique to find out the accuracy of our method. Totally 6888 machine print words and 5801 hand written words used for classification. we divided the data into 10 groups and then use 9 groups as training set and 1 group as test set. The same procedure is repeated 10 times with different sets of training and testing data. The classification efficiency of each cross fold validation and the corresponding overall average efficiency showed in Table II.

We compared our approach with [2]. Our approach achieved an overall accuracy of 98.6 % on IAM dataset compared to 94.5% of [2]. The main differences observed in our method with respect to [2] are the classifier and the set of features used. The classifier used in [2] is defined by using marginal (threshold) values of each feature. These threshold values of machine print and handwritten text words are overlapping to each other, which causes decrease in efficiency. Also the features used in [2] are based on intensity only, but not structural properties of the text. Our approach uses a combination of structural and intensity features and SVM classifier for classification, for better performance.

V. CONCLUSION

This paper presents a set of features for printed and handwritten text detection. We used a new set of com-
TABLE II
Classification accuracy of handwritten and printed word separation for IAM dataset

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Cross fold Numbers</th>
<th>Classification Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cross fold 1</td>
<td>97.91</td>
</tr>
<tr>
<td>2</td>
<td>Cross fold 2</td>
<td>99.84</td>
</tr>
<tr>
<td>3</td>
<td>Cross fold 3</td>
<td>99.86</td>
</tr>
<tr>
<td>4</td>
<td>Cross fold 4</td>
<td>98.42</td>
</tr>
<tr>
<td>5</td>
<td>Cross fold 5</td>
<td>99.12</td>
</tr>
<tr>
<td>6</td>
<td>Cross fold 6</td>
<td>97.23</td>
</tr>
<tr>
<td>7</td>
<td>Cross fold 7</td>
<td>98.53</td>
</tr>
<tr>
<td>8</td>
<td>Cross fold 8</td>
<td>99.78</td>
</tr>
<tr>
<td>9</td>
<td>Cross fold 9</td>
<td>98.25</td>
</tr>
<tr>
<td>10</td>
<td>Cross fold 10</td>
<td>97.78</td>
</tr>
<tr>
<td>11</td>
<td>Total Average</td>
<td>98.6</td>
</tr>
</tbody>
</table>

Combination of intensity and structural features with SVM classifier and achieved satisfactory results with a standard IAM dataset[13]. We plan to use some new properties like smoothness in curved shapes and space between connected components. In future we are planning to extend this work for different Indian scripts text classification.

REFERENCES


