An Efficient OCR System based on the Regional Feature using the ASVM as Classifier

Maninder Kaur  
Research Scholar, ECE Department, Rayat and Bahra University, Mohali

Ms. Manjeet Kaur  
Assistant Professor, ECE Department, Rayat and Bahra University, Mohali

ABSTRACT

In Image Processing, sometimes due to poor handwriting, the writer left some gap between diacritics and character or between diacritics and header line due to which small text blocks gets created which leads to improper text line segmentation and hence leads to wrong results and overlapping. As a result accuracy of the algorithm degrades. In proposed work Adaptive SVM will be used to improve accuracy of the system.

INTRODUCTION:

Optical character recognition which is also commonly known as optical character reader is the process of converting the mechanical and electronic images into the handwritten, printed text etc. OCR is a course by which focused software is used to change the skimmed pictures of manuscript to electronic text so that digitized data can be examined, indexed and recovered. The OCR are basically design to settled and improved the multiple real world applications such as mining data from business documents, checks, passports, invoices, bank statements, computerized receipts, business cards, mail, printouts of static-data, or any suitable documentation. OCR is mostly used in the area of the computer vision, artificial intelligence, and pattern recognition. Handwriting text recognition (HTR) can be defined as the ability of a computer to transform handwritten input represented in its spatial form of graphical marks into equivalent symbolic representation as ASCII text. Usually, this handwritten input comes from sources such as paper documents, photographs or electronic pens and touch-screens.

Fig.1 Handwritten document Convert into Text

Approaches for learning Optical character recognition:- The following are the approaches for learning Optical character recognition.

➢ Histogram Approach:- This method is based on the pixel histogram in which a Y-histogram forecast is achieved which results in text line position and to divide the line into different areas a threshold is applied.
OCR involved various steps to read the characters from a scanned Image. In proposed research, a model has been built for handwritten images. The system extracts the characters from handwritten images and writes into text file.

**Flow Chart OF OCR Model**

1. Pre-processing
2. Segmentation
3. Normalization
4. Feature extraction
5. Classification
6. Post-processing

**Fig.3 Steps of OCR System**

**Data Acquisition:** Most Important initial phase in OCR is to gather the image from either device sensor like PDA or tablets in case on online recognition or getting the images containing characters directly for offline recognition. The image should have a specific format such as JPEG, BMP etc.

**Pre Processing:** The goal of pre-processing is to simplify the pattern recognition problem without missing any vital information. It reduces the noises and inconsistent data. It enhances and prepares it for the next steps.

**Segmentation:** Segmentation is an integral part of any text based recognition system. It assures efficiency of classification and recognition. Accuracy of character recognition heavily depends upon segmentation phase.

**Normalization:** The results of segmentation process provides isolated characters which are ready to pass through feature extraction stage, thus the isolated characters are reduced to a specific size depending on the methods used. The segmentation process essentially renders the image in the form of m*n matrix.

**Feature Extraction:** Feature extraction is the process of extracting the relevant features from the text images.
objects/alphabets to form a feature vectors. These feature vectors is then used by classifiers to recognize the input unit with target output unit. Feature extraction methods are based on 3 types of features:

- Statistical
- Structural
- Global transformations and moments

**Classification:** The results Classification is the last stage where we train the neural net using the feature vectors obtained during feature extraction method against the required targets.

**Post Processing:** The goal of post processing is the incorporation of context and shape information in all the stages of OCR systems is necessary for meaningful improvements in recognition rates.

**Feature extraction:** The following is the feature matching and classification algorithm for matching the extracted plant disease image with the different images of same plant, which are taken at different times, from different viewpoints, or by different sensors.

**RESULT AND DISCUSSION:**

<table>
<thead>
<tr>
<th>No.</th>
<th>Images</th>
<th>Real Text</th>
<th>Recognized Text</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>The electrical resistance of an electrical conductor is a measure of the difficulty to pass an electric current through that conductor.</td>
<td>The electrical resistance of an electrical conductor is a measure of the difficulty to pass an electric current through that conductor.</td>
<td>96.52</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Difficult roads often lead to beautiful destinations</td>
<td>Difficult roads often lead to beautiful destinations</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>It is a field of research in pattern recognition, artificial intelligence and machine vision.</td>
<td>It is a field of research in pattern recognition, artificial intelligence and machine vision.</td>
<td>97.50</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Morning walk is a good exercise. An early-riser can be a regular morning walker. The benefit of morning walk are manifold.</td>
<td>Morning walk is a gooo exercise. An early-riser can be a regular morning walker. The benefit of morning walk are manifold.</td>
<td>95.09</td>
</tr>
<tr>
<td>Page</td>
<td>Image</td>
<td>Text</td>
<td>Accuracy</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-------</td>
<td>------</td>
<td>----------</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td><img src="image1.png" alt="Image" /></td>
<td>The future is what will happen in the time after the present. Its arrival is considered inevitable due to the existence of time and the laws of physics.</td>
<td>97.56</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td><img src="image2.png" alt="Image" /></td>
<td>Teamwork makes the dreamwork</td>
<td>80.00</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td><img src="image3.png" alt="Image" /></td>
<td>Neural networks can be used, if we have a suitable dataset for training and learning purposes. Dataset are one of the most important things when constructing new neural network.</td>
<td>98.62</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td><img src="image4.png" alt="Image" /></td>
<td>Good things take time</td>
<td>83.34</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td><img src="image5.png" alt="Image" /></td>
<td>Older OCR systems match these images against stored bitmaps based on specific fonts.</td>
<td>98.61</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td><img src="image6.png" alt="Image" /></td>
<td>Social media are computer-mediated technologies that facilitate the creation and sharing of information ideas, career interests, and other forms of expression via virtual communities and networks.</td>
<td>97.60</td>
<td></td>
</tr>
</tbody>
</table>

**Overall Accuracy of Proposed work** | **94.48**
In above figure, the accuracy of proposed work is represented in the form of graph. In graph, X-axis denotes the number of samples which are included in the proposed work for the testing and Y-axis denotes the accuracy of proposed work in percentage. From the above graph, it has been observed that the average percentage of accuracy is more than 94% with handwritten images.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>ANN</th>
<th>SVM</th>
<th>ASVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td>86</td>
<td>93</td>
<td>94</td>
</tr>
</tbody>
</table>

Fig.4 Accuracy of proposed work

Fig.5 Accuracy comparison between Classifier
In above figure, the accuracy comparison between artificial neural network (ANN), support vector machine (SVM) and adaptive support vector machine (ASVM) is represented in the form of bar graph. From the figure we observe the accuracy of proposed character recognition system with ASVM is better than ANN and SVM classifier due to the best training.

**Conclusion:** Due to overlapping and touching of characters, there remains no significant gap between the text lines and hence two or more text lines comes in a same text block which leads to wrong results. The main focus in this research project is to experiment deeply with, and find alternative solutions to the image segmentation and character recognition problems within the Overlapped Character Recognition. In the existing work, SVM classifies is applied but it has less accuracy. So in future, Adaptive SVM will be applied to improve better accuracy of the system.

**Future work:** In future, we can use the artificial neural network along with the optimization algorithm to achieve the better results by minimizing the more noisy data from the images for the character recognition system. The combination of the artificial neural network as classifier instead of SVM with optimization technique the precision of character recognition will have to increase and the rate of noise will decreases.

**REFERENCES**

16) Ping Wang, Wei Zhang, "Research and Realization of Improved Pattern Matching in


