**Optical Character Recognition**

**Sai Krishna Amsanpally Shen Yang**

[**Skaf48@mail.missouri.edu**](mailto:Skaf48@mail.missouri.edu) [**Syg84@mail.missouri.edu**](mailto:Syg84@mail.missouri.edu)

**Abstract**

Optical Character Recognition (OCR) is one of the widely known and studied technique for automatic identification/conversion of text from graphical (typed or hand written) or scenic images into machine readable form. It achieves this by classifying patterns corresponding to characters’ present in the images. Automatic identification is applied in different areas like speech recognition for assistance in booking reservations, Radio frequency for identifying cars, Bar codes for identifying products etc. In many applications, information is not available in human readable form. OCR helps to have information present in human readable form and machine can detect the information. Some of the important applications of OCR are for information entry from printed data records like passport documents, invoices, bank statements etc. It is the common method for digitizing printed texts to make them electronically editable, searchable, stored more compactly. This project aims to extract the text of a scanned document and make it available in a word processor.

**1 Introduction**

Text recognition in both images as well as videos has gained significant attention from technological community in recent times with advancements in mobile phones technology. There are mobile applications available today which translate text into other languages in real time that stimulated interest in this domain. Text is very expressive form of communication and often included in images for more effective communication. OCR performs text recognition for both hand written and printed characters and the results directly depend on quality of input images. Images with lower quality and degraded data are still posing challenges to this well studied problem. There is still room for improvement with text detection and character recognition rates falling below 80% and 60% respectively. Applying advanced computer vision and pattern recognition techniques will help improve the results.

**2 Background**

Text is basically into two classes, graphic text and scenic text. Graphic text refers to text that is overlaid on images or objects graphically while scenic text refers to the one appears in natural scenes, sign boards, clothes and most probably includes hand written characters. Significant amount of research is done for OCR of graphical text while numerous methods are introduced lately for recognizing text in scenic images. The primary goal here is to check if there is any text in the image or not and recognize it. These basic set of tasks that does this job are Text localization and verification, Text extraction (focuses on Text binarization). OCR has important applications. Automatic sign recognition on boards and translation will help people with language barriers. A personal text to speech device can assist visually impaired or blind to recognize signs for groceries, product labels, help using ATM machines etc. Industrial automation heavily uses OCR technology.

Numerous factors like scene complexity, uneven lighting, blurring and degradation, aspect ratios, distortion, multiple fonts and multilingual environments influence the quality of text recognition results. Below figure shows one of the methodologies followed in OCR.

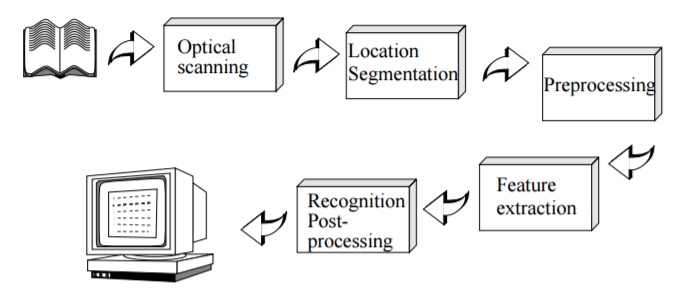


Fig. 1. Components of an OCR-system

The localization step crudely classifies the components and groups them into text regions. In verification phase, these regions are further classified as text or non-text regions. Segmentation separates the characters into exclusive outlines of image blocks so that they are converted to characters in recognition step. Segmentation is done by isolating each connected component. It’s easy to implement but poses challenges if the text is in cursive and have connections, or if they are split.

Preprocessing images will remove any noise present n images and smoothen them which includes filling and thinning. Filling eliminates gaps and holes while thinning reduces width if the lines. Preprocessing normalizes the images to obtain characters with uniform size, slant and orientation.

Feature extraction captures the most discriminant characteristics of symbols. The most basic way of telling a character is by the image itself. Other ways of feature extraction are transformations and Structural analysis.

**4 Proposed methods**

Template matching and correlation is the most basic way of recognizing a character. No features are extracted in this approach. The input image in the form of matrix is compared with set of prototype characters of each possible class. The class of the matching template with highest correlation is assigned to input image.

1. (b)

Fig 2.a Template Fig 2.b Input image after character segmentation

In feature based techniques, measurements are calculated and extracted from training characters and compared to test character. The description that matches closely recognizes the character. The feature is a vector of numbers and represents a symbol. Distribution of points, zoning, moments, crossings and distance are other ways of extracting features.

**KNN:**

**Decision Trees:**

A binary classification decision tree for multiclass classification and an ensemble of classifiers are used to train and recognize text in images. Trees in general do a split in training data based on the values of the available features. The split at each node is based on the feature that gives the maximum information gain. At the end, each leaf node corresponds to a class label. The algorithm can handle binary or multiclass classification problems.

**5 Data:**

We used 74k dataset which is one of the popular datasets to work with in OCR research. The character set is limited to English alphabets and Arabic numerals. The dataset consists of:

1. 62 classes (0-9, A-Z, a-z)
2. 7705 characters obtained from natural images
3. 3410 hand drawn characters using a tablet PC
4. 62992 synthesized characters from computer fonts

**6. Results:**

Template matching approach:

|  |  |  |
| --- | --- | --- |
| Test image | Text recognized | Accuracy |
| C:\Users\saikr\AppData\Local\Microsoft\Windows\INetCache\Content.Word\krishna.png | kclgood | 14.29 |
| C:\Users\saikr\AppData\Local\Microsoft\Windows\INetCache\Content.Word\test3.png | mollotmlsgscoockgohcodcdctoccccozogtmoo | 41.03 |

Table 1. Template matching method results

KNN

|  |  |  |  |
| --- | --- | --- | --- |
| #Neighbors >>>>  Training Set | 1 | 3 | 128 |
| Written characters | 0.6774 | 0.7258 | 0.8226 |
| Characters in different fonts | 0.7419 | 0.7258 | 0.8548 |

Table 1. Confusion matrix for KNN results

Decision Trees:

The 3410 images in handwritten characters’ dataset is divided coarsely in the ratio of 90:10. Training dataset has 3100 images while test set has 310 images.

|  |  |
| --- | --- |
| Algorithm | Prediction Accuracy |
| Single classification tree: | 98.08 |
| Ensemble of classifiers | 99.06 |

Table 2. Classification accuracy details using trees

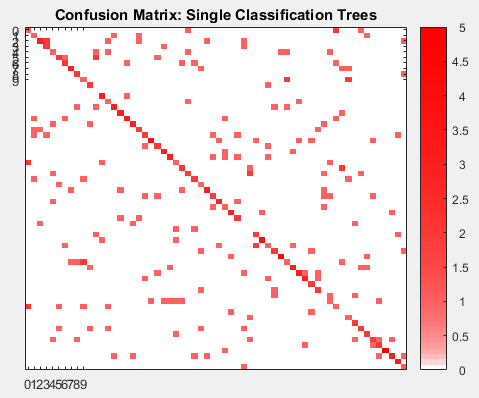
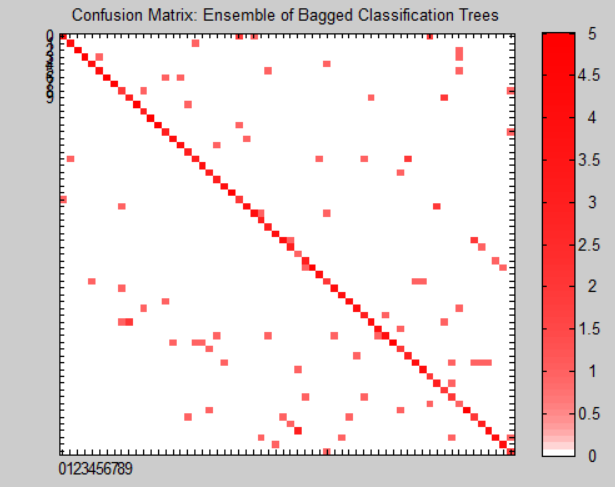
 

Fig 3. Classification results using ensemble of classifiers

**Conclusion and Future work:**

The results have improved significantly with decision trees over KNN and template matching approaches. The results look great because test and train sets are part of same original hand written characters dataset. The algorithm should be fed all the 74k training images to test on a new unknown image for text recognition to know the strength of algorithm in classifying images and recognizing text.

For future work, feature extraction should be done which are designed based on different representations of the characters, such as solid binary characters, character contours, skeletons (thinned characters) or gray-level sub images of each individual character. Same dataset can be evaluated with different classification algorithms like SVM and neural network and compare the results.

**References**

[1] <http://www.ee.surrey.ac.uk/CVSSP/demos/chars74k/>

[2] <https://en.wikipedia.org/wiki/Optical_character_recognition>

[3] <https://vision.cornell.edu/se3/wp-content/uploads/2014/09/wang_iccv2011.pdf>

[4] <http://ieeexplore.ieee.org/abstract/document/6945320/>