OCR WITH MACHINE LEARNING

Deepak Jannarapu **III year ECE (13116021)**

Nallagatla Manikanta **III year ECE (13116045)**

Keetha Indraneel Varma **III year EE (13115062)**

Palakonda Shiva Prasad **III year ECE (13116047)**

Koneti Sarat Kumar **III year EE (13115065)**

Abstract:

Handwritten recognition has been a difficult problem to solve for a computer. Various handwritings and different shapes of the alphabets has been a major difficulty in recognising the alphabets. The computer will require doing complex processes such as image processing, recognising lines, segmentation.

To improve accuracy we have applied a machine learning technique in our project. We have used multi support vector machine for training and testing the data. A new method is introduced for extracting features of alphabets. We have used fifty datasets each containing 26 alphabets(capital letters) and 50 datasets containing 10 numeric handwritten characters for training. Our project uses MATLAB environment for image processing and for applying machine learning techniques.

The required image of the form is given as input to the system. Handwritten characters are filtered out using RGB values of the individual pixels. And machine divides the image into segments of characters. These characters are sent for testing. Characters are detected accordingly and are dynamically stored in the excel file.

Acknowledgement:

This project was carried out under the guidance of Electronics Section, Indian Institute of Technology. In completing our project we had to take the help and guidance of some respected persons, who deserve our greatest gratitude. We would like to show our gratitude to Mr Kamal Singh Gotyan for giving us good guidelines for this project. We would also like to show our gratitude to Padmanabh Pande for his ideas and suggestions for this project throughout numerous consultations. We would also like to expand our deepest gratitude to Rahul Ratan Mirdha, Gaurav Waghmare and Advait Vaidya who have directly and indirectly guided us in completing this project.

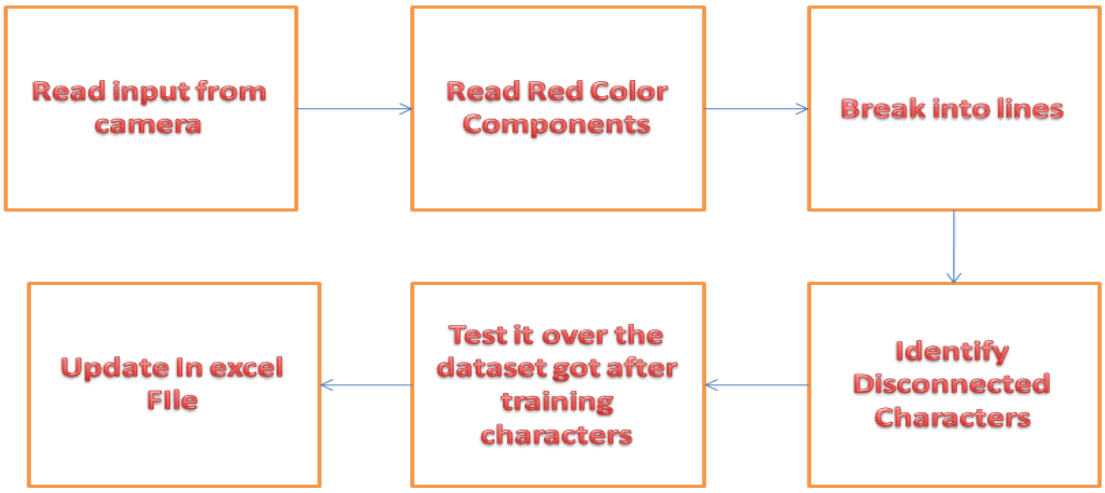
Many people, especially our classmates and team members itself, have made valuable comments and suggestions on this project which gave us inspiration to improve our project. We thank all the people who had directly and indirectly helped to complete our project successfully.

Introduction:

The current capacity to translate paper documents quickly and accurately into machine readable form using optical character recognition technology augments the opportunities in document searching and storing, as well as the automated document processing.
OCR stands for Optical character recognition. It is a technology that recognizes and captures alphanumeric characters on a computer at a high speed. It provides complete form processing and documents capture solution. It is one such a system that allows us to scan printed, hand written text (numerals, text) and convert scanned image into a computer process able format. In this project we convert the user handwritten data into excel data sheet. OCR reduces the data entry time and increases its accuracy when compared to the use of manual data entry operators. The system stores data in a database thus facilitating data analysis. It reduces the number of data entry personnel.

The process of OCR involves several steps including:
Segmentation
Feature extraction
classification

Program Overview:

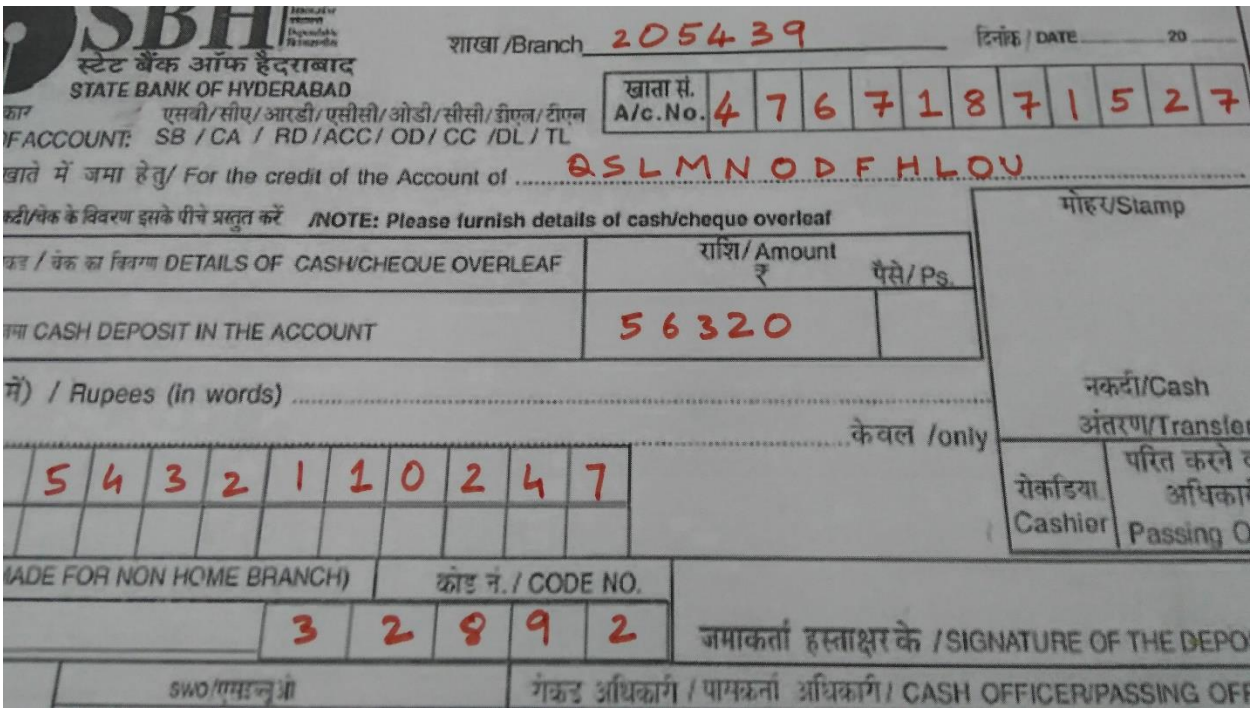


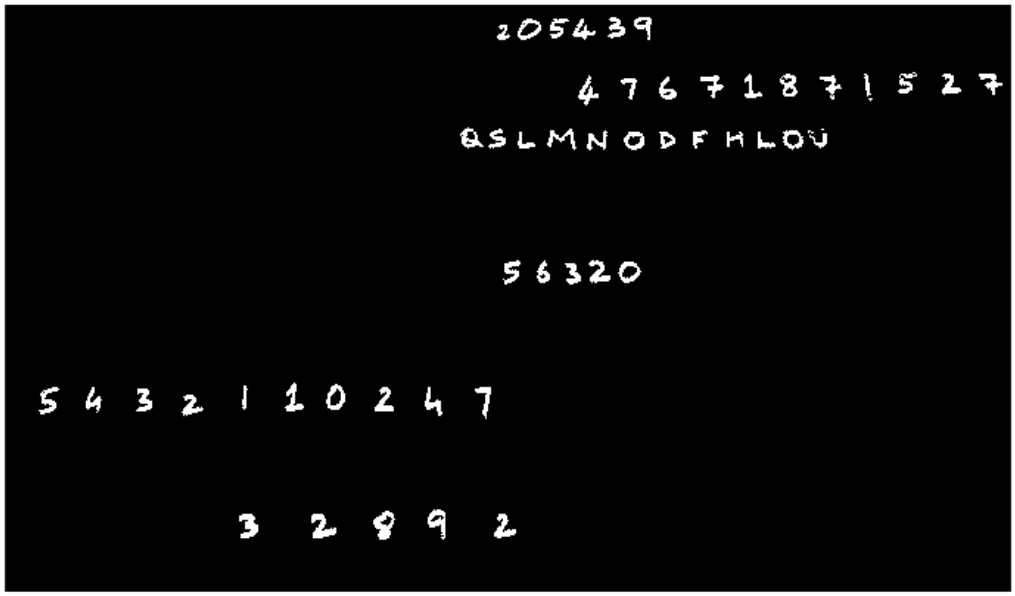
Algorithm:

By using webcam we can take image of the form in which details are filled in red color. Convert the given RGB image into HSI image we can easily separate the red color pixels using hue component. By using red color thresholding we can stop the all the pixels which are not in red color. Then we can remove the noise using the median filter and use dilation and erosion to remove small dots and connect the near disconnected small group of pixels. By this method, we can get only the red pixels which we need without any noise. Once we have the details we can break the image into lines by using image segmentations with the help of the spaces present between the lines. We can process each line and label all the connected components into characters.

By using the machine learning algorithm Multi SVM, we have trained the handwritten set of characters to their targets. Once we have each character we can test it with trained multi SVM and get the results. If the results are not correct then it means the training set need to be updated. So we take input from the user for correcting the error characters and update the training set.

Results:





data - Excel

File | Home | Insert | Page Layout | Formulas | Data | Review | View | Tell me what you want to do...

Cut | Copy | Format Painter | Paste | Clipboard | Calibri | 11 | Font | Wrap Text | Merge & Center | Alignment | General | Number

F10 | 32892

	A	B	C	D	E	F	G
1	Branch	Account Number	Name	Amount	Mobile number	Branch Code	
2	IIT	123456789	DEEPAK	1E+11	999999999	123	
3	1069	13321467895	INDRANEEL	48679	9897654321	101069	
4	1004	13121461041	SNDRANZEL	40619	9897614321	101000	
5	1069	13321467895	IZDRANEEL	48679	9890654321	1060	
6	1032	98976628650	IITROORKEE	17398	9876543210	1.03E+13	
7	1	99040795357	REDLIGAHT	3575	4097553799	14343	
8	1	99040795357	REDLIGHT	3575	4097553799	14343	
9	100209	47671871527	GSLMNODFBLZVB	6110	5002110147	12192	
10	205439	47671871527	QSLMNODFHLOU	56320	5432110247	32892	
11							
12							

Conclusion:

Our system is designed for recognizing characters and numbers in a bank form in real time. For the implementation of this task we used Multi SVM as a tool. The tests we did went smoothly and we had no problems, except for the fact that we had to use large datasets. We have implemented this project in a way that we can train the machine indefinitely. Efforts have been made to reduce the errors and increase the accuracy. The application is implemented using MATLAB in Windows environment using high quality camera. Our application can also be used for automatic data detection from forms used in hospitals, train services etc.

Future Work:

From 1950s OCR is an active area of research. Our project can be useful in many ways considering the detection of multiple types of handwritings. Many techniques for recognition of Offline English Handwritten Characters have been suggested. But still an efficient OCR for the recognition of handwritten letters does not exist. Few steps have been taken for Handwritten and Hand printed (which is a constrained handwritten) English letter recognition. There is a large scope in future.

- It can be used as font independent optical character recognition tool.
- There is a heavy demand to reduce cumbersome work involved in converting the handwritten forms to digital forms. The work involved in this process can be reduced.
- Extensive features can also be added to software for translating into another language or for converting text into speech.
- Data which is detected can also be stored into cloud which can be used anywhere and anytime.

References:

1. Ivan Dervisevic, Machine Learning Techniques for Optical Character Recognition, 2006.

2. G. Smith, Optical Character Recognition, CSIRO Manufacturing and Infrastructure Technology, Australia, 2003.

3..http://www.cvisiontech.com/resources/ocr-primer/ocr-neural-networks-and-other-machine-learning-techniques.html