HANDWRITTEN CHARACTER RECOGNITION

This project report is submitted to

Silicon Institute of Technology, Bhubaneswar

in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology

in

Computer Science and Engineering

Submitted by

Shubham Gantayat (190310360)

Rajat Pati (190310313)

Ayush Mohanty (190310265)

Dibyajyoti Das (190310194)

Sujan Kumar Patro (190310418)

Under the Esteemed Supervision of

Prof. Rabindra Kumar Dalei

Prof. Jiten Mohanty

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SILICON INSTITUTE OF TECHNOLOGY

SILICON HILLS, BHUBANESWAR – 751024, ODISHA, INDIA

June, 2022

ACKNOWLEDGEMENTS

We are highly indebted to **Silicon Institute of Technology, Bhubaneswar** for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project. We would like to express our gratitude towards Prof. **Rabindra Kumar Dalei** for guiding us throughout the project. We thank all participants for their positive support and guidance. Our heartfelt thanks to

Mr. Jiten Mohanty for assisting us throughout our project. We believe we will enrollin more such events in the coming future.

Subham Gantayat
Rajat Pati
Ayush Mohanty
Dibyajyoti Das
Sujan Kumar Patro

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SILICON INSTITUTE OF TECHNOLOGY
BHUBANESWAR – 751024

ABSTRACT

With the knowledge of current data about a particular subject, machine learning tries to extract hidden information that lies in the data. By applying some mathematical functions and concepts to extract hidden information, machine learning can be achieved and we can predict output for unknown data. Pattern recognition is one of the main applications of ML. Patterns are usually recognized with the help of a large image data-set. Handwriting recognition is an application of pattern recognition through image.

By using such concepts, we can train computers to read letters and numbers belonging to any language present in an image. There exist several methods by which we can recognize hand-written characters. We will be discussing some of the methods in this paper.

Keywords: *Machine learning*

LIST OF ABBREVIATIONS

Abbreviation	Description
NP	NumPy
PD	Pandas
LR	Logistic Regression
ACC	Accuracy
RE	Recall
PRE	Precision
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
RFC	Random Forest Classifier
PLT	PyPlot

LIST OF FIGURES

LIST OF TABLES

CONTENTS

CHAPTER 1

INTRODUCTION

Handwritten character recognition (HCR) is the process of conversion of handwritten text into machine readable form. The major problem in the handwritten character recognition (HCR) system is the variation of the handwriting styles, which can be completely different for different writers. The objective of the handwritten character recognition system is to implement user friendly computer assisted character representation that will allow successful extraction of characters from handwritten documents and to digitize and translate the handwritten text into machine readable text.

Handwritten character Recognition system is divided into two categories

- On-line character recognition.

  It is a system in which recognition is performed when characters are under creation.

- Off-line character recognition.

  It is a system in which first handwritten documents are generated, scanned, stored in a computer and then they are recognized.

Handwritten Character Recognition System consists of following stages:

1) Pre-processing.
2) Segmentation.
3) Feature extraction.
4) Training and recognition.
5) Post processing.

1.1. BACKGROUND

In the field of Machine Learning, recognition of objects has become the most sought one. Some of the examples of object recognition are Face recognition, Handwriting recognition, Disease detection etc. All these things can happen through a large set of image dataset. These image dataset will contain both positive and negative data regarding that domain. This helps the algorithm to classify the unknown data in better ways. Handwriting recognition is a new

technology that will be useful in this 21st century. It can act as base functionality for the birth of new requirements. For example, a blind man cannot read a newspaper unless braille format exists. In this case we can train the algorithm to recognize characters in the newspaper, store them as text and convert the text to speech. This can help a lot of blind people to ease their daily work. The second application of handwriting recognition could be language translation. In this case when a person is dealing with non-native language, he can just take an image of a document and send it to the hand write recognition algorithm. This algorithm can recognize the characters in the image and convert them to text. Then the text can be converted to the desired language of choice. brought to you by CORE provided by International Journal of Computer (IJC - Global Society of Scientific Research and... International Journal of Computer (IJC) (2020) Volume 38, No 1, pp 93-101 94 * Corresponding author. One more application of handwriting recognition would be, processing of large sets of paper documents like answer scripts. With the help of hand-write recognition and AI, the answer scripts can be evaluated without human involvement. For all above mentioned scenarios, handwriting recognition acts as a base case to be resolved. Handwriting recognition is one of the types of Optical Character Recognition (OCR). OCR is identification of text, which may be printed or hand-written. In OCR, the document is captured via camera as an image and canbe converted to desired formats like PDFs. Then the file is fed to the algorithm for character recognition. This can drastically reduce human involvement in certain scenarios.

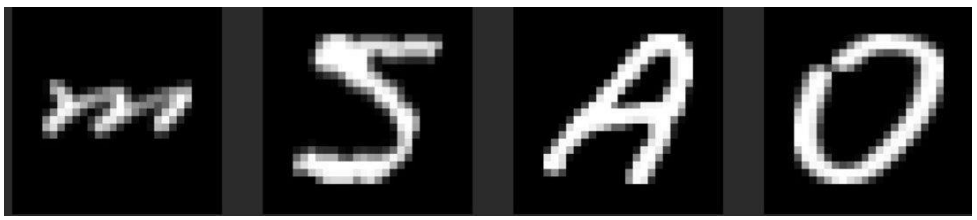
1.2. PROBLEM STATEMENT

In this project, we have to classify 26 English alphabets using machine learning techniques.

1.3. PROPOSED METHOD

1.3.1. Method to Improve the Performance of Discovery

Dataset:



It has been shown that Support Vector Machines (SVMs) can be applied to image and hand-written character recognition [4]. SVMs are effective in high dimensional spaces, hence it

makes sense to use SVMs for this study given the high dimensionality of our input space, i.e., 784 features. However, SVMs don't perform well in large datasets as the training time becomes cubic in the size of the dataset. This could be an issue as our dataset contains 42,000 samples which is quite large. To deal with this issue, we will adopt a technique proposed by a study conducted at the University of California, Berkeley, which is to train a support vector machine on the collection of nearest neighbors in a solution they called "SVM-KNN" [2]. Training an SVM on the entire data set is slow and the extension of SVM to multiple classes is not as natural as Nearest Neighbor (NN). However, in the neighborhood of a small number of examples and a small number of classes, SVMs often perform better than other classification methods.

We use NN as an initial pruning stage and perform SVM on the smaller but more relevant set of examples that require careful discrimination. This approach reflects the way humans perform coarse categorization: when presented with an image, human observers can answer coarse queries such as presence or absence of an animal in as little as 150ms, and of course, can tell what animal it is given enough time [6]. This process of a quick categorization, followed by successive finerbut slower discrimination was the inspiration behind the "SVM-KNN" technique.

1.3.2. Evaluation Parameter – Classification Report

A Classification report is used to measure the quality of predictions from a classification algorithm. How many predictions are True and how many are False. More specifically, True Positives, False Positives, True negatives and False Negatives are used to predict the metrics of a classification report

1.4. PROJECT ORGANIZATION

Chapter 1 titled, "*Introduction", presents the* general overview of the concept of image segmentation, techniques and applications involving it. It also contains the objective and motivation behind the work.

Chapter 2 titled, "Literature review*" presents the renowned* works earlier performed by well-known personalities in the area of image processing. This chapter furthermore contains the positives and negatives of the existing works done.

Chapter 3 named as, "*Proposed Method",*

Chapter 4 titled, "Experimental Results*"*

Chapter 5 titled, "*Conclusion", is the* summary of the complete work carried out with a miniature part given to the society.

SUMMARY

In this chapter we discussed about:

- The basic concept and goals of Logistic Regression.
- Various applications where image segmentation plays a vital part.
- Parameters for evaluating the results like Precision, Accuracy, F1-score and recall from confusion matrix.
- Motivation, objective and domain of this work.

CHAPTER 2

LITERATURE REVIEW

Handwritten character recognition is a field of research in artificial intelligence, computer vision, and pattern recognition. A computer performing handwriting recognition is said to be able to acquire and detect characters in paper documents, pictures, touch-screen devices and other sources and convert them into machine-encoded form. Its application is found in optical character recognition, transcription of handwritten documents into digital documents and more advanced intelligent character recognition systems.

Handwritten character recognition can be thought of as a subset of the image recognition problem.

Basically, the algorithm takes an image (image of a handwritten digit) as an input and outputs the likelihood that the image belongs to different classes (the machine-encoded alphabets, A–Z). In this blog post, I will elaborate on my approach to solving this problem with a combination of machine learning techniques.

CHAPTER 3

METHODOLOGY

PROPOSED SYSTEMS

Preprocessing Methods:

Stratified Random Split:

Provides train/test indices to split data in train/test sets. This cross-validation object is a merge of Stratified Fold and Shuffle Split, which returns stratified randomized folds. The folds are made by preserving the percentage of samples for each class.

Standard Scaler:

Standardize features by removing the mean and scaling to unit variance.

Models Used:

Logistic Regression:

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable.

Random Forest Classifier:

Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

Performance was better in Random Forest for our use case.

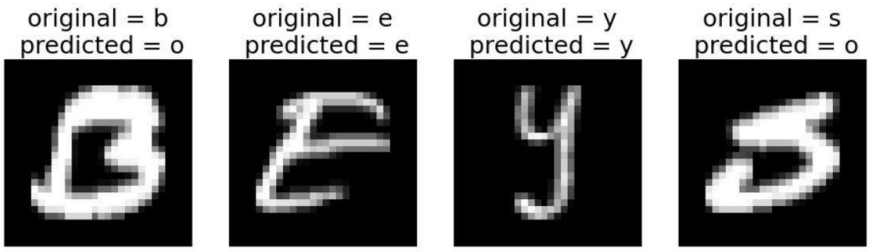
CHAPTER 4

Experimental Results

Logistic Regression performance:

	precision	recall	f1-score	support
0	0.86	0.84	0.85	3534
1	0.77	0.86	0.81	1958
2	0.89	0.90	0.89	5823
3	0.78	0.83	0.80	2359
4	0.80	0.83	0.82	2764
5	0.82	0.89	0.86	268
6	0.78	0.84	0.81	1328
7	0.72	0.77	0.75	1673
8	0.78	0.94	0.85	234
9	0.80	0.80	0.80	2137
10	0.76	0.76	0.76	1398
11	0.94	0.91	0.92	3019
12	0.88	0.87	0.88	3127
13	0.82	0.81	0.81	4854
14	0.96	0.93	0.95	14818
15	0.92	0.90	0.91	4994
16	0.77	0.83	0.80	1336
17	0.79	0.81	0.80	2852
18	0.94	0.94	0.94	12094
19	0.94	0.92	0.93	5753
20	0.89	0.88	0.88	7344
21	0.88	0.90	0.89	1022
22	0.80	0.82	0.81	2633
23	0.82	0.83	0.82	1556
24	0.85	0.83	0.84	2784
25	0.85	0.89	0.87	1451
accuracy			0.88	93113
macro avg	0.84	0.86	0.85	93113
weighted avg	0.88	0.88	0.88	93113

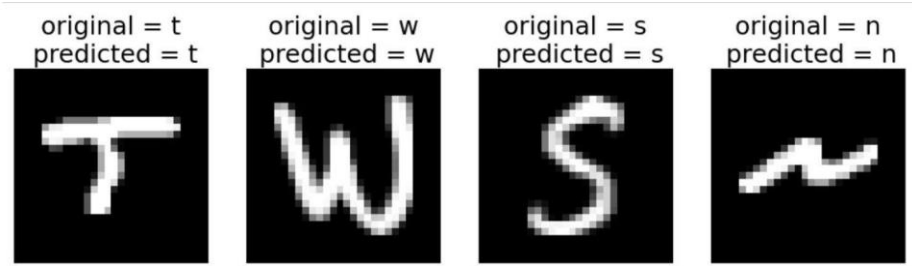
Output of Logistic Regression:



Random Forest Performance

	precision	recall	f1-score	support
0	0.99	0.97	0.98	3553
1	0.96	0.98	0.97	2117
2	1.00	0.99	0.99	5886
3	0.96	0.98	0.97	2487
4	0.98	0.99	0.98	2835
5	0.93	1.00	0.96	270
6	0.96	0.99	0.97	1394
7	0.97	0.98	0.97	1782
8	0.95	1.00	0.98	267
9	0.97	0.99	0.98	2080
10	0.96	0.97	0.97	1386
11	0.99	0.99	0.99	2909
12	0.97	0.99	0.98	3031
13	0.99	0.98	0.98	4810
14	1.00	0.99	0.99	14614
15	0.99	0.99	0.99	4874
16	0.94	0.98	0.96	1396
17	0.98	0.98	0.98	2873
18	1.00	0.99	0.99	12166
19	1.00	0.99	0.99	5649
20	1.00	0.99	0.99	7309
21	0.99	0.99	0.99	1041
22	0.98	0.99	0.98	2653
23	0.97	0.99	0.98	1535
24	0.98	0.99	0.98	2696
25	0.98	0.99	0.99	1500
accuracy			0.99	93113
macro avg	0.97	0.99	0.98	93113
weighted avg	0.99	0.99	0.99	93113

Output images by Random Forest Classifier:



CHAPTER 5

CONCLUSION

As the Random Forest Classifier (accuracy = 98.6%) performs better than the Logistic Regression Classifier (accuracy = 87.9%) on the test data set, we can conclude that it is more suitable for our use case.

REFERENCES

- **www.kaggle.com/sachinpatel21/az-handwritten-alphabets-in-csv-format/**
- **www.scikit-learn.org/**

APPENDIX – A

LIST OF FORMULAS USED

1. Accuracy(acc)%= ((TP+TN)/(TP+TN+FP+FN)) *100
2. Recall (Re)%=(TP/(TP+FN)) *100
3. Precision (Pre)%=((TN)/(TN+FP)) *100
4. F1-Score=2*(Precision)*(Recall)/ (Precision Recall)

(Backside Hard Cover of the Project Report)

.

Department of Computer Science and Engineering
Silicon Institute of Technology,
Silicon Hills, Bhubaneswar –751024,
Odisha, India