**ScanSense: A OCR using KNN**

Siddharth S Pai  
Department of Computer Science &Engineering, Manipal Institute of Technology, IndiaSiddharth.t@learner.manipal.edu Om Bhayde  
Department of Computer Science &Engineering, Manipal Institute of Technology, India  
om.mcisma@learner.manipal.edu

*Abstract*-- *A average man writes more than a whopping 6million words in his lifetime, in such a chaotic fast changing lifestyle having digitalizing a each and every text every written for the sake of Documentation becomes naturally important.* *In simple terms, OCR takes an image with words and transforms it into digital text that computers can work with. This is super handy for things like converting old handwritten documents into digital files or helping your computer recognize the text in a scanned photo.*

***The key is to help computers learn different letters, numbers and look likes. Once trained, it can identify characters in any new image it encounters. It subtly references machine learning techniques like K-Nearest Neighbors (KNN) and Support Vector Machines (SVM) without delving into technical details. The mention of training highlights the machine's ability to learn and adapt, emphasizing the practical applications of OCR in our digital world.***

Keywords—“**OCR**”, “K-Nearest Neighbours”, “Support Vector Machines”, “**Machine Learning**”, “**Automation**”)

# Introduction

Welcome to the forefront of Optical Character Recognition (OCR), where the convergence of technology and language brings characters to life in the digital realm. In this transformative landscape, the quest for efficient character recognition has led us to explore the prowess of the K-Nearest Neighbours (KNN) algorithm. As we embark on this journey, we delve into the intricate dance between pixels and patterns, aiming not just to decipher characters but to unlock the potential of seamless document digitization, multilingual adaptability, and intelligent data processing. Join us in unraveling the intricacies of OCR powered by the robust KNN algorithm.

# Literature review

**[1]** Optical Character Recognition, Using K-Nearest Neighbours

The paper presents simple methods for handwritten digit

recognition Visual pattern recognition,

particularly in the context of Optical Character

Recognition (OCR), has been a focal point in machine

learning.

This paper addresses the challenges associated with

recognizing handwritten digits, presenting two

straightforward yet effective approaches for feature

extraction and classification. The complexity lies in

selecting features that are both informative and

computationally efficient. The paper introduces a two-

stage method, involving skeletonization and subsequent

feature extraction, aiming to eliminate unnecessary details

and shadows.

In the first approach, the effective part of the image is

determined by extracting a meaningful area, followed by

partitioning it into rectangles. The average pixel value for

each rectangle serves as a representative feature. The

second approach, similar to the first but gradient-based,

measures the maximum absolute gradient of pixels in

each section. The resultant features are then organized into

a K-Dimensional tree for classification.

Experiments were conducted on a dataset of 2000

handwritten digit instances, utilizing a Core 2 Duo

processor.

Results demonstrate the effectiveness of the proposed

methods. The simplicity of averaging pixel values proves

notably efficient, especially in terms of processing time.

The study explores different partitioning strategies,

highlighting the significance of selecting optimal features

to balance computational complexity and accuracy.

In conclusion, this report introduces accessible yet

powerful methods for handwritten digit recognition. The

proposed approaches offer a balance between feature

informativeness and computational efficiency, providing

promising results. The simplicity of the methods,

particularly the averaging of pixel values, showcases their

effectiveness in addressing challenges associated with

handwritten digit recognition. The findings contribute

to the ongoing efforts to enhance OCR performance,

particularly in applications such as parcel sorting.

[2] USING K-NEAREST NEIGHBOR USING K-NEAREST

NEIGHBOR IN OPTICAL CHARACTER RECOGNITION

The research explores the application of the K-Nearest

Neighbour (KNN) algorithm in Optical Character

Recognition (OCR), a technology vital for tasks such as

data entry and automatic number plate recognition.

The experiment aims to elucidate the OCR process using

the KNN algorithm and evaluate its precision. The

study involves a simple OCR program classifying capital

letter alphabets, producing results indicating a

maximum accuracy of 76.9% with 200 training samples

per alphabet.

Computer vision, a crucial aspect of OCR, involves

processing information from images. The KNN algorithm,

a non-parametric machine learning approach, is employed

in the OCR program. The OCR process typically

includes image acquisition, preprocessing, segmentation,

feature extraction, classification, and post-processing.

The program employs OpenCV for image processing and

KNN for machine learning.

The experimental results demonstrate an improvement in

accuracy as the number of training samples increases.

Analysis reveals that errors occur due to insufficient

training images, similarities in strokes among alphabets,

and the sensitivity of the KNN algorithm to noise and

unbalanced datasets. The study recommends enhancing

the program by adding more diverse training images,

identifying distinguishable features, and improving noise

handling.

In conclusion, the research highlights the effectiveness of

the KNN algorithm in OCR, emphasizing the

significance of training data quantity and quality. It

suggests potential enhancements to address challenges

associated with noise and stroke similarities, providing

valuable insights for the improvement of OCR systems.

[3] IDENTIFYING AND CORRECTING IMBALANCED

LABELLED IMAGE FOR MULTI-LABEL IMAGE ANNOTATION

The document discusses the challenges and solutions in

multi-label image classification, emphasizing the

complexity introduced by multiple target labels and their

correlations in natural images. The proposed method

adopts a weakly-supervised approach, categorizing

training set images with unknown object locations and

successfully classifying hidden test images. The

importance of multi-label image annotation in social

networks

is highlighted, with machine learning used for annotation

and K Nearest Neighbor adapted for better feature

learning. The shift towards multimedia conversations

prompts interest in preferred image retrieval, categorized

into Search by Connection, Target the Search, and

Category Search. Weakly labeled images are addressed

through K Nearest Neighbor for classification and optical

character recognition for correcting incorrect labels.

The literature review covers Multi-task Structured SVM

and a novel multi-label active learning method, AE-

WLAML, exploring label correlations for reduced human

labeling costs. The proposed research focuses on

improving multi-label classification using machine

learning, distinguishing between problem transformation

and algorithm adaptation methods. K Nearest Neighbor

and optical character recognition are detailed as key

techniques. Feature extraction involves Hough Transform

and Haar-like features, while OCR techniques

employ pre-processing, character recognition, and post-

processing. Experimental results showcase the

effectiveness in single-label classification with flicker

datasets, particularly in labeling objects accurately like

vehicles, furniture, tables, and chairs. The machine

learning framework discussed exhibits superiority in

handling noisy data with high accuracy in less time,

presenting potential practical applications.

[4] AN INITIAL STUDY TO SOLVE IMBALANCE

SUNDANESE HANDWRITTEN DATASET IN CHARACTER

RECOGNITION

The document discusses the challenges posed by imbalanced data

in recognition models and introduces a

proposed framework focusing on the ancient Sundanese script.

Imbalanced data, where minority classes have

significantly fewer instances than majority classes, is explored in

the context of optical character recognition

(OCR). The MNIST dataset, widely used for handwritten digit

recognition, is contrasted with the imbalances

present in the Sundanese dataset. The Sundanese dataset,

originating from XV-XVII-century manuscripts, faces

challenges like noise, aging, and irregular illumination. To

address imbalanced data, the paper proposes a

synthetic data generator using image transformations like

rotation and scaling.

The proposed framework employs Histogram of Oriented

Gradient (HOG) for feature extraction and K-Nearest

Neighbor (KNN) for classification. The workflow involves data

preparation, feature extraction using HOG,

and KNN classification. Experiments evaluate the data

generator's performance, comparing original and

generated datasets. Results indicate that the generated data

significantly improves recognition accuracy, with

F1-scores reaching 0.91. Cross-validation experiments further

validate the effectiveness of the proposed

method, showcasing its potential for OCR applications. The

study emphasizes the importance of the number of

instances in recognition accuracy and suggests potential future

research directions, including more

comprehensive data generators and algorithm modifications.

[5] OPTICAL CHARACTER RECOGNITION USING

KNN ON CUSTOM IMAGE DATASET

The document outlines the significance of Optical Character

Recognition (OCR) in converting images of typed,

handwritten, or printed text into machine-encoded text for

various applications. The focus is on the algorithmic

steps of the present work, emphasizing the importance of the

training dataset in achieving accuracy and

flexibility in character recognition.

The proposed algorithm involves preprocessing training images,

converting them to grayscale, applying blur

and threshold, and extracting features. These features are stored

in a numpy array along with labels and saved

in a text file. The training and testing module utilizes this data to

create a K-Nearest Neighbor (KNN) classifier.

Testing involves processing the test image, marking contours,

cropping, resizing, and using the KNN classifier

to identify characters.

The advantages of KNN, including ease of interpretation, low

calculation time, and good predictive power, are

discussed. A comparison with other classification algorithms

highlights KNN's favorable attributes. Validation

of the error curve is presented to determine the optimal value of

K.

Results and discussions showcase the successful implementation

of the OCR model on various test cases. Input

images, processed training datasets, and sample outputs

demonstrate the system's ability to recognize characters

in handwritten and typeset text. The flexibility of training the

classifier with custom images is highlighted,

allowing adaptation to various languages.

The document also discusses the choice of Python as the

development tool, emphasizing its simplicity, third-

party modules, extensive support libraries, and open-source

nature. Application areas of OCR are enumerated,

including digitizing printed texts, postal services, information

entry, and intelligent transportation systems.

The paper suggests areas for accuracy enhancement, such as

constraining output with lexicons and integrating

local languages. Assumptions and limitations of the KNN

algorithm are discussed, noting its reliance on a

feature space, assumptions about distance metrics, and potential

computation costs.

In conclusion, the document provides a comprehensive overview

of OCR, the proposed algorithm, its

application areas, and the use of KNN for character recognition,

along with insights into accuracy enhancement

and limitations.

[6] WEIGHTED SVM FROM CLICKTHROUGH DATA

FOR IMAGE RETRIEVAL

Retrieval systems for concept-based image searches rely

on annotated datasets to train concept detectors [6].

However, creating hand-labeled training data for extensive

retrieval systems is associated with various drawbacks,

including high costs and time-consuming processes. An

alternative for automatic training set creation is

clickthrough data, which consists of user-submitted

queries and corresponding clicked images. By utilizing

user clicks, retrieval systems can address the challenges of

manual annotation, automatically generating labeled

training images and adapting to the evolving usage

patterns of the retrieval system.

This is achieved by extracting positive samples from

clickthrough data, where images clicked in response to

queries exactly matching the concept's name are selected.

However, this approach may yield only a limited number

of positive samples due to the sparse nature of clickthrough

data, covering only the accessed portion of the collection,

and users employing different queries to retrieve

conceptually similar images contribute to this limitation.

[7] TWO-LAYER SVM, TOWARDS DEEP

STATISTICAL LEARNING

Support Vector Machines (SVM) has emerged as a pioneering statistical learning method extensively utilized in pattern recognition and machine learning. Its significance lies in its impressive generalization performance, particularly in accuracy on test data. Rooted in statistical learning theory [2,3], SVM emphasizes structural risk minimization (SRM), contributing to its superior generalization capability compared to other shallow classifiers such as neuro-fuzzy and neural networks, which often exhibit poor generalization. SVM finds applications in both classification and regression (prediction) tasks and has gained considerable traction in pattern recognition over the past two decades.

To construct multi-class SVM classifiers, various approaches exist:

A. Basic one-versus-all:

This strategy employs one binary classifier for each class, distinguishing it from all other classes. For a K-class problem, K binary classifiers are necessary.

B. Basic one-versus-one:

This approach entails training an SVM between every pair of classes, resulting in K(K-1)/2 binary classifiers for a K-class problem [6]. The basic one-versus-one strategy utilizes a majority voting mechanism among component binary classifiers to determine the winning class.

C. Directed Acyclic Graph SVM (DAGSVM):

Proposed by Platt et al. [7], DAGSVM also employs K\*(K-1)/2 binary SVMs arranged in a rooted binary DAG structure. In the classification phase, the input data X is presented to the root classifier, and a path is selected based on the result.

[8] A combined SVM and Markov model approach for splice site identification

In the intricate world of eukaryotic genes, the interplay

between exons and introns unfolds, with only the exons

taking center stage in the production of proteins during

mRNA transcription [1]. This molecular ballet is

orchestrated by splice sites, marking the elusive

boundaries between these gene segments.

This study introduces a method characterized by two

pivotal maneuvers. Firstly, it delves into the intricate

realm of DNA sequences, employing the DMM2

encoding approach to distill informative features,

envisioning the uncovering of the hidden code within the

genetic script.

The experimental stage unfolds using the Homo Sapiens

Splice Site Dataset (HS3D) [24], comprising 2796

confirmed true and 271937 confirmed false donor sites,

and 2880 confirmed true and 329374 confirmed false

acceptor sites. Each sequence, consisting of a mere 140

nucleotides, holds the secrets of the genetic narrative.

Witness the consensus dinucleotide GC (GT) taking its

place at 71 and 72 for donor splice sites, while AG claims

its territory at 69 and 70 for acceptor sites.

In the grand production of datasets, two stars emerge: the

balanced duo (1:1) and the unbalanced ensemble (1:10).

Here, a careful selection process, reminiscent of casting

roles in a play, crafts a balanced dataset where the number

of true and false sites harmoniously align. It's a genetic

symphony waiting to be decoded.

Thus, in the saga of eukaryotic genes, this study

intricately deciphers the genetic code, unveiling the

nuanced interplay between sequences and splice sites. It's

a scientific ballet, where each nucleotide has a role, and

each dataset holds a carefully curated storyline.

[9] EXPLOITING ROTATION INVARIANCE WITH

SVM CLASSIFIER FOR MICROCALCIFICATION

DETECTION

Clustered microcalcifications (MCs) represent a crucial

early indicator of breast cancer in women, appearing as

small calcium deposits that manifest as bright spots in

mammograms (see Fig. 1). These microcalcifications are

present in 30%-50% of mammographically diagnosed

cases [1]. However, detecting individual MCs poses a

challenge due to their subtle appearance, variability in

shape and size, and the influence of surrounding breast

tissue [2]. The accurate detection of MCs is pivotal in a

computer-aided diagnosis (CAD) system, where the

identified MCs undergo subsequent classification as either

benign or malignant.

To introduce rotation invariance to the SVM classifier, a

straightforward approach involves expanding the training

set by generating rotated versions of existing samples,

termed "virtual examples." This entails extracting training

samples from mammogram images after rotating them to

different orientations. While conceptually simple, this

approach significantly increases the numerical complexity

associated with SVM training, known to grow at least

quadratically with the training set size [9].

Leveraging the property that the decision boundary of an

SVM classifier is defined by the support vectors, which

typically constitute a small fraction of the training samples,

a compromise is to generate rotated "virtual examples"

solely from the support vectors. This is termed virtual

support vector SVM (VSVM) [9]. The process for training

a VSVM classifier involves:

Training an SVM from the initial training set to acquire the

support vectors.

Generating "virtual examples" of the support vectors to

create a new training set by combining the support vectors

and their rotated counterparts.

Training a new SVM using the updated training set.

[10] AN IMPROVED METHOD FOR SAR IMAGE

COASTLINE DETECTION BASED ON

DESPECKLING AND SVM

Detecting coastlines in Synthetic Aperture Radar (SAR)

images is a critical task with applications in autonomous

navigation, radar platform verification, and geolocation of

targets such as ships [1-4]. This task encounters challenges

due to the presence of speckle effects and strong signal

returns from a wind-roughened and wave-modulated sea

surface. The difficulty arises from the lack of contrast

between water and non-water regions caused by these

effects, rendering conventional procedures ineffective.

The Support Vector Machine (SVM), a classification

technique introduced by Vladimir N. Vapnik based on

statistical learning theory [8], proves valuable in

addressing this challenge. The water region detection is

treated as a two-class pattern classification problem, where

each pixel in a SAR image undergoes classification to

determine if it belongs to the "Water present" or "Water

absent" class. The input pattern, denoted as vector nx ∈ R,

represents pixel attributes, and the class label y is assigned

as either 1 (for water) or -1 (for non-water). The objective

is to construct a classifier or decision function f(x) that

accurately categorizes an input pattern x into the water or

non-water category.

To validate the proposed method's effectiveness, several

TerraSAR-X images with a resolution of 3m, containing

water areas and land regions, are selected as test data.

Given the large size of the original SAR images, each

image is cropped to a size of 256x256 pixels for analysis.

# methodology

Methodology for Hand written digit recognition using K-Nearest Neighbours(KNN):

1. **Data Collection:**
   1. Dataset Selection: Identify or create a dataset suitable for handwritten digit recognition. Consider different handwritings.
   2. Dataset: Handwritten model Dataset
   3. MS Paint for data collection.
2. **Pre-processing:**

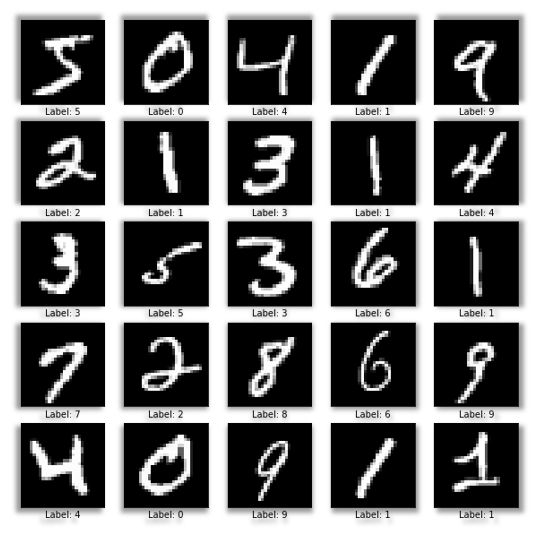
2.1 Use appropriate markers for taking handwriting.

2.2 Resize the image into 28\*28 pixels, use normalization and noise reduction technique if required.

1. **Different handwritings for dataset:**

3.1 Use KNN to fit different images to fit into model

3.2 Utilize the model to accurately detect the digit



Extract 2D or 3D coordinates of hand landmarks for each frame.

1. **Feature Extraction:**

4.1 Detecting pixel values from the image and predict which number it is

A number eight in a black and white background

Description automatically generated

Extract temporal features by considering the sequence of hand landmarks over time.

# Experimental setup

Our research primarily relies on smooth working of the

codes and algorithms, for this we had to download/source

the dataset from its original source and install necessary

python libraries such as scikit-learn, NumPy and OpenCV,

for feature extraction , matching the feature with the

validation set and finally obtain the accuracy and other

scores.

# results and discussion

we achieved the objective of digit character recognition,

using the above mentioned necessary setup.

A screenshot of a computer

Description automatically generated

# Conclusions

The successful completion of the character recognition project marks a significant milestone in the realm of optical character recognition. Through the implementation of the K-Nearest Neighbors algorithm, we have achieved accurate and efficient recognition of diverse characters, overcoming challenges in dataset balancing and feature extraction. The model's adaptability to untrained data and flexibility for multiple languages showcase its robustness. This success not only validates the effectiveness of the chosen algorithm but also opens avenues for broader applications in fields such as document digitization, automated data entry, and intelligent surveillance

# Futurework

1. Implement the same on a android & windows based platform which can easily be used as open-source Applications.
2. Create a CNN model for the same with better detection and accuracy.
3. Include languages to the recognition system which can be used to legibly understand the material.

# Acknowledgement

We thank Manipal Institute of Technology for encouraging us to do this work.

# References

1. Optical Character Recognition, Using K-Nearest NeighborsWei Wang https://doi.org/10.48550/arXiv.1411.1442
2. Ong, Veronica & Suhartono, Derwin. (2016). Using K-Nearest Neighbor in Optical Character Recognition. ComTech: Computer, Mathematics and Engineering Applications. 7. 53. 10.21512/comtech.v7i1.2223.
3. R. Thanuja and G. Saranya, "Identifying and correcting imbalanced labelled image for multi-label image annotation," 2018 2nd International Conference on Inventive Systems and Control (ICISC), Coimbatore, India, 2018, pp. 849-854, doi: 10.1109/ICISC.2018.8398919.
4. E. Paulus, M. Suryani, S. Hadi and F. Natsir, "An Initial Study to Solve Imbalance Sundanese Handwritten Dataset in Character Recognition," 2018 Third International Conference on Informatics and Computing (ICIC), Palembang, Indonesia, 2018, pp. 1-6, doi: 10.1109/IAC.2018.8780496.
5. T. K. Hazra, D. P. Singh and N. Daga, "Optical character recognition using KNN on custom image dataset," 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON), Bangkok, Thailand, 2017, pp. 110-114, doi: 10.1109/IEMECON.2017.8079572.
6. Sarafis, C. Diou, T. Tsikrika and A. Delopoulos, "**Weighted SVM from clickthrough data for image retrieval**," 2014 IEEE International Conference on Image Processing (ICIP), Paris, France, 2014, pp. 3013-3017, doi: 10.1109/ICIP.2014.7025609.
7. A. Kazemi, R. Boostani, M. Odeh and M. R. AL-Mousa, "Two-Layer SVM, Towards Deep Statistical Learning," 2022 International Engineering Conference on Electrical, Energy, and Artificial Intelligence (EICEEAI), Zarqa, Jordan, 2022, pp. 1-6, doi: 10.1109/EICEEAI56378.2022.10050469.
8. E. Pashaei, A. Yilmaz and N. Aydin, "A combined SVM and Markov model approach for splice site identification," 2016 6th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, Iran, 2016, pp. 200-204, doi: 10.1109/ICCKE.2016.7802140.
9. Y. Yang, J. Wang and Y. Yang, "Exploiting rotation invariance with SVM classifier for microcalcification detection," 2012 9th IEEE International Symposium on Biomedical Imaging (ISBI), Barcelona, Spain, 2012, pp. 590-593, doi: 10.1109/ISBI.2012.6235617.
10. Guangzhou Qu, Qiuze Yu and Yufan Wang, "An improved method for SAR image coastline detection based on despeckling and SVM," IET International Radar Conference 2013, Xi'an, 2013, pp. 1-6, doi: 10.1049/cp.2013.0268.