

TEXT EXTRACTION FROM IMAGES USING MACHINE LEARNING

A PROJECT REPORT

Submitted by

ANURAG KUMAR	(7th Sem)	2101020505
TANIYA ANSHU	(7th Sem)	2101020507
SAJAN KUMAR	(7th Sem)	2101020509

In partial fulfilment for the award of the degree of

**BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE & ENGINEERING**



**C.V. RAMAN GLOBAL UNIVERSITY
BHUBNESWAR- ODISHA -752054**

NOVEMBER 2024



C.V. RAMAN GLOBAL UNIVERSITY
BHUBANESWAR-ODISHA-752054

BONAFIDE CERTIFICATE

Certified that this project report "**Text Extraction from Images using Machine Learning**" is bonafide work submitted by **ANURAG KUMAR, 7th Semester, Registration No. 2101020505, TANIYA ANSHU, 7th Semester, Registration No. 2101020507, SAJAN KUMAR, 7th Semester, Registration No. 2101020509, CGU-Odisha, Bhubaneswar** who carried out the project under my supervision.

Dr. MONALISA MISHRA
HEAD OF THE DEPARTMENT
Department of Computer Science &
Engineering

Dr. MAMATA P. WAGH
SUPERVISOR
Associate Professor, Department of
Computer Science & Engineering



**C.V. RAMAN GLOBAL UNIVERSITY
BHUBANESWAR-ODISHA-752054**

CERTIFICATE OF APPROVAL

This is to certify that we have examined the project entitled "**Text Extraction from Images using Machine Learning**" is bonafide work submitted by **ANURAG KUMAR, 7th Semester, Registration No. 2101020505, TANIYA ANSHU, 7th Semester, Registration No. 2101020507, SAJAN KUMAR, 7th Semester, Registration No. 2101020509,** CGU-Odisha, Bhubaneswar.

We hereby accord our approval of it as a major project work carried out and presented in a manner required for its acceptance for the partial fulfillment for **Bachelor Degree of Computer Science and Engineering** for which it has been submitted. This approval does not necessarily endorse or accept every statement made, opinion expressed, or conclusions drawn as recorded in this major project, it only signifies the acceptance of the major project for the purpose it has been submitted.

Dr. MAMATA P. WAGH
SUPERVISOR
Associate Professor, Department of
Computer Science & Engineering

DECLARATION

The team has successfully completed the “**Text Extraction from Images using Machine Learning**” project, adhering to the guidelines and instructions provided by our instructor. All submitted work is original, with appropriate citations and acknowledgement of external sources. The work has not been previously submitted for any other course or assignment. The project was completed to the best of our abilities and in a timely manner, and we are confident that it meets the required standards and expectations set by our instructor. This research was conducted with honesty, integrity, and transparency, and has not been submitted for any academic degree or examination. All sources used in the report have been appropriately cited, and any errors or omissions are the responsibility of the author.

Anurag Kumar 2101020505

Taniya Anshu 2101020507

Sajan Kumar 2101020509

Bhubaneswar – 752054

19-11-2024

ACKNOWLEDGMENT

We would like to articulate our deep gratitude to our project guide Dr. Mamata P. Wagh, Associate Professor, Department of Computer Science and Engineering, who has always been source of motivation and firm support for carrying out the project. We would also like to convey our sincerest gratitude and indebtedness to all other faculty members and staff of Department of Computer Science and Engineering, who bestowed their great effort and guidance at appropriate times without it would have been very difficult on our project work. An assemblage of this nature could never have been attempted with our reference to and inspiration from the works of others whose details are mentioned in the references section. We acknowledge our indebtedness to all of them. Further, we would like to express our feeling towards our parents and God who directly or indirectly encouraged and motivated us during Assertion.

Anurag Kumar 2101020505

Taniya Anshu 2101020507

Sajan Kumar 2101020509

ABSTRACT

Text extraction from images, also known as Optical Character Recognition (OCR), plays a critical role in digitizing and processing visual data. However, extracting text from complex backgrounds, blurred images, or varied lighting conditions remains a significant challenge. This report presents a machine learning-based approach to improve text extraction accuracy using Support Vector Machines (SVM) for segmentation and classification, integrated with Tesseract OCR for character recognition.

The proposed model leverages SVM to detect and classify text regions from images based on features like texture and edge patterns. After segmenting the text regions, Tesseract OCR extracts the actual text, providing a robust solution to issues commonly faced in traditional OCR, such as handling multiple orientations, fonts, and noisy backgrounds. The model was trained and evaluated using the IIIT5K Word dataset, which contains diverse real-world scene images.

Experimental results show that the model performs reliably in extracting text from high-quality images and relatively simple backgrounds. However, challenges remain with blurred or noisy images, where character segmentation affects the accuracy of the final output. Future work aims to extend the model's capabilities to handle multi language texts, enhance segmentation for low-quality images, and improve real-time processing.

This project highlights the potential of combining SVM-based segmentation with OCR to develop more accurate and efficient text extraction systems, particularly for applications involving natural scene images and document digitization.

TABLE OF CONTENTS

DESCRIPTION	PAGE NUMBER
BONAFIDE CERTIFICATE	i
CERTIFICATE OF APPROVAL	ii
DECLARATION	iii
ACKNOWLEDGEMENT	iv
ABSTRACT	v
LIST OF FIGURES	vii
LIST OF TABLES	viii
1. INTRODUCTION	1-3
1.1 : Background	1-2
1.2 : Problem Statement	2
1.3 : Objectives	3
2. LITERATURE SURVEY	4-6
3. METHODOLOGY	7-14
3.1 : Data Collection and Preparation	7
3.2 : Workflow	7-8
3.3 : Model Architecture	8-9
3.4 : Model Impression	9
3.5 : Machine Learning Techniques	10-12
3.6 : Deep Learning Techniques	13-14
4. RESULTS AND DISCUSSION	15-19
4.1 : Analysis of Extracted Output	15-16
4.2 : Calculation of Evaluation Metrics	17-18
4.3 : Comparative Analysis	18-19
5. CONCLUSION	20
REFERENCES	21

LIST OF FIGURES

FIGURE	TITLE	PAGE NUMBER
3.1	Workflow of the model	8
3.2	Flow Chart of the model with an example	9
4.1	Bar Chart of Confusion Matrix	18
4.2	ROC of Accuracy	19

LIST OF TABLES

TABLE	TITLE	PAGE NUMBER
3.1	Comparison of Models on the basis of Advantages, Limitations, and Typical Accuracy.	11
3.2	Comparison of Techniques	13
4.1	Extracted output of the input images under SVM	15
4.2	Extracted output of the input images under SVM	16
4.3	Evaluation of Model Accuracy	18

Chapter 1

INTRODUCTION

Text extraction from images, also known as Optical Character Recognition (OCR), involves detecting and extracting text from a variety of image sources such as street signs, posters, documents, or handwritten notes. This process is a critical component of many applications, including document digitization, automated form reading, and real-time text translation from street signs or product labels.

1.1 Background

Text extraction from images is a critical task in applications like document digitization and automated translations. However, it remains challenging due to factors such as intricate backgrounds, variability in font styles, sensor limitations, resolution issues, and varying light conditions. These complexities often overwhelm OCR systems, resulting in low accuracy and poor generalization across diverse scenarios [1]. Moreover, Basic level approaches don't cope with certain problems such as distorted text, or overlapping fonts or complex document layout. Carrying out such approaches through a sequential process seems quite impossible in real situations which leads to poor applicability of the majority of traditional approaches. Also, SVM (Support Vector Machines) excels in machine learning, particularly for image classification, by effectively handling high-dimensional data. Its core functionality lies in margin maximization, ensuring robust generalization by identifying the optimal hyperplane that separates classes with the widest possible margin [2]. This project employs methods like SVM for segmentation and image text classification, chosen for its precision and scalability, ensuring efficient handling of text extraction tasks while addressing challenges like overlapping characters and complex backgrounds effectively. Thus, using OCR systems and integrating them with these models enhanced by the addition of preprocessing techniques, our main goal is to build a framework which will overcome the most popular issues of traditional text extraction approaches.

The integration of machine learning techniques has emerged as a powerful approach to address these limitations. Support Vector Machines (SVM), a supervised learning model, have shown immense potential in enhancing text extraction processes. By leveraging features like edges, textures, and colour patterns, SVM can effectively distinguish text from non-text regions, even in challenging visual environments. SVM's ability to classify segmented text regions based on high-dimensional feature spaces makes it particularly suitable for pre-processing steps in OCR systems, such as text region identification and character classification.

This project focuses on building a hybrid framework that combines the strengths of OCR systems with SVM-based classification techniques. By incorporating advanced pre-processing methods, the system ensures improved segmentation of text regions, enabling OCR to achieve higher recognition accuracy. The goal is to create a scalable solution that not only addresses the shortcomings of traditional OCR methods but also excels in handling real-world complexities such as blurred text, varied fonts, and noisy surroundings. This integration represents a step forward in enhancing text extraction technologies and broadening their applicability in diverse domains.

1.2 Problem Statement

It is a challenging task to extract text from images, especially when complex backgrounds, varied fonts, changing lighting conditions, and low-resolution inputs exist. The variability involved is difficult for conventional OCR systems to capture, which often results in low accuracy and poor generalization. Furthermore, existing solutions do not address problems such as distorted text or overlap of text and complex document layouts. The inability to robustly handle these scenarios makes traditional approaches less applicable in real world situations. This project attempts to overcome these challenges by implementing advanced models such as SVM, with higher accuracy and scalability, to segment and classify text from images. We are therefore interested in integrating these models with preprocessing techniques and OCRs to develop a more robust framework for overcoming the constraints presented by traditional text extraction methods.

1.3 Objectives

The main goal of this project is to create a machine learning based system that is capable of robust and effective text extraction from images. The system aims to meet the critical requirements of scene text recognition such as noisy image backgrounds, poor contrast of text in images and text appearing in multiple orientations. More specifically, the project is structured around the following tasks.

Accurate Segmentation: In this regard automatic detection and extraction of the text regions in images where the background scene is variously complicated is important.

Enhanced Recognition of Text: This involves SVM based text classification and Tesseract OCR based text extraction from the segmented regions to increase recognition effectiveness.

Robustness and Stability: Focus is also on reliable functioning when images contain text written in various fonts, angles as well as conditions under control.

Real World Use Cases: The objective is also to develop a system that has real life application, for example, in capturing text from documents, reading through forms automatically, and translating text on the image in real time.

Chapter 2

LITERATURE SURVEY

As discussed, earlier text recognition from images is still an active research in the field of pattern recognition. To address the issues related to text recognition many researchers have proposed different technologies, each approach or technology tries to address the issues in different way. In forthcoming section, we present a detailed survey of approaches proposed to handle the issues related to text recognition.

Text detection and recognition in complex and cluttered environments has gained significant attention due to its applications in document analysis, scene text extraction, and image processing. Various models and techniques have been proposed to address the challenges posed by diverse text styles, orientations, and backgrounds. This section reviews several notable approaches in the field, highlighting both their strengths and limitations, particularly in handling complex scenes.

Optical Character Recognition (OCR) has advanced from simple document scanning to complex text recognition in natural scenes. Yuming He [1] highlighted its role in converting images to machine-readable text, addressing challenges such as handwriting styles and complex backgrounds. Traditional methods like SVM were effective for printed text but struggled with scene variability. Modern deep learning-based OCR systems significantly enhance functionality.

Deepa et al. [2] explored the application of Support Vector Machines (SVMs) and Convolutional Neural Networks (CNNs) in image classification. SVMs were highlighted for their ability to handle high-dimensional data efficiently, leveraging margin maximization to achieve robust generalization. The flexibility of kernel functions, including linear kernels, enables SVMs to perform both linear and non-linear classification, making them valuable in domains like medical imaging and document classification. CNNs, on the other hand, excel in hierarchical feature extraction through convolution and pooling layers, automating filter optimization and capturing abstract features, which significantly enhances performance in tasks like object recognition.

Ghai et al. [3] emphasized the role of K-means clustering in text detection, highlighting its simplicity, efficiency, and adaptability. As an unsupervised learning algorithm, K-means effectively handles heterogeneous datasets, making it suitable for detecting text with varying fonts, sizes, orientations, and backgrounds. The study applied K-means to segment high-frequency wavelet features extracted from decomposed image. By iteratively minimizing intra-cluster variance and maximizing inter-cluster separation, the algorithm optimized cluster centers using statistical measures like mean and standard deviation of wavelet coefficients. This approach proved effective in distinguishing text from non-text regions in complex image datasets.

Tang et al. [4] introduced Cascaded Convolutional Neural Networks (CNNs) for text detection and segmentation, emphasizing their hierarchical processing capabilities. This architecture refines outputs progressively, enabling effective extraction of local features like edges and textures, as well as global context. By structuring tasks into extraction, refinement, and classification phases, cascaded CNNs enhance precision while minimizing false positives. The multi-layer refinement strategy resolves ambiguities and consolidates overlapping text regions, ensuring high accuracy and reliability even in complex scenes with various text styles, orientations, and sizes. This systematic approach makes Cascaded CNNs highly effective for robust text detection tasks.

Rong et al. [5] proposed a novel approach for text segmentation in cluttered scenes by integrating visual and linguistic features through referring expressions. The model utilizes both visual data and natural language descriptions, enabling effective segmentation, and was evaluated on the COCO-CharRef dataset. This method significantly enhances contextual text segmentation by combining multimodal inputs, improving accuracy in complex scenarios. However, the reliance on annotated datasets limits the model's adaptability to new environments, highlighting a need for more generalized solutions.

Yousef et al. [6] proposed a CNN-based model for handwriting recognition that demonstrates high efficiency in handling both short and long lines of text across various handwriting styles and sizes. The model effectively recognizes unconstrained text with minimal data, making it versatile for diverse applications. However, its performance

declines when applied to paragraphs or multiple lines of text without additional segmentation algorithms, highlighting a limitation in processing extensive textual content.

Surana et al. [7] reviewed several methodologies for character recognition, emphasizing their evolution and effectiveness. Back Propagation Networks (BPN) were among the earliest techniques, leveraging the Radon Transform combined with Back Propagation Neural Networks for image-to-binary conversion. This method segmented images into sub-images, converting individual characters into binary format, which proved instrumental for scanned text recognition. Artificial Neural Networks (ANNs) advanced these efforts by introducing a three-layer architecture optimized using the Scaled Conjugate Gradient method. This approach surpassed traditional backpropagation techniques, achieving a recognition accuracy of 95%, thereby establishing ANNs as a reliable model for text classification.

Chapter 3

METHODOLOGY

Text extraction from images, also known as Optical Character Recognition (OCR), involves detecting and extracting text from a variety of image sources such as street signs, posters, documents, or handwritten notes. This process is a critical component of many applications, including document digitization, automated form reading, and real-time text translation from street signs or product labels.

3.1 Data Collection and Preparation

The model is trained and tested on the IIIT5K Word dataset, which is widely used for scene text recognition tasks. This dataset consists of 5,000 images of words collected from real-world scenes such as street signs, billboards, and various natural images. The images in the dataset vary in terms of text size, font style, orientation, and background complexity, providing a comprehensive set of challenges for training a robust text extraction model. The dataset helps the model learn to generalize across different conditions, ensuring that it performs well not only in controlled environments but also in real-world scenarios where text extraction is crucial.

The primary objective of this project is to build a machine learning-based solution that accurately segments and extracts text from images, overcoming the challenges of lighting variation, noise, and complex backgrounds. This is achieved through:

- Using Support Vector Machines (SVM) for image segmentation and classification of text regions.
- Implementing Tesseract OCR for actual text extraction from the segmented regions.

3.2 Workflow

The Implementation Sequence outlines the systematic steps involved in processing an image for tasks such as text recognition. Each stage plays a critical role in ensuring accurate and efficient results. Below is a brief description of each step in the sequence as shown in Fig. 3.1.

- **Image Acquisition:** This is the initial step where images are captured using a device like a camera or scanner. It involves obtaining a digital image of the object or text to be processed.
- **Preprocessing:** This stage prepares the image for further analysis by enhancing its quality. It includes steps like noise removal, contrast enhancement, resizing, and converting the image to a binary or grayscale format.
- **Character Segmentation:** In this step, the processed image is divided into individual characters. This is crucial for recognizing text, as each character is isolated for analysis.
- **Feature Extraction:** This step involves identifying unique and relevant features from the segmented characters. These features are used as input for classification algorithms, enabling recognition or further analysis.

Implementation Sequence

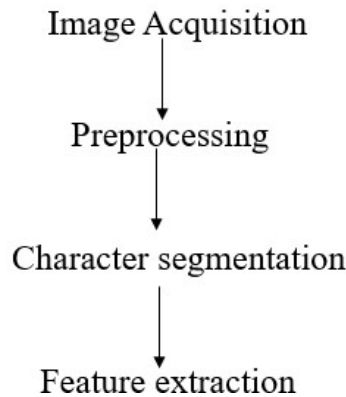


Fig 3.1: Workflow of the model

3.3 Model Architecture

The proposed solution aims to tackle the challenges of text extraction from images by combining Support Vector Machine (SVM) for classifying text and non-text regions, with Tesseract Optical Character Recognition (OCR) for recognizing and extracting text from the identified regions. The process begins with image preprocessing to enhance the quality of the input by reducing noise, improving contrast, and highlighting edges, which helps in better feature extraction for text detection. SVM is then applied to segment

the image by analyzing features like texture, colour patterns, and edges, effectively classifying which parts of the image are likely to contain text. Once the text regions are detected, Tesseract OCR is used to extract and convert the segmented text into readable and editable formats as shown in Fig 3.2. Tesseract is an open-source OCR engine known for its ability to recognize text across different fonts, orientations, and languages, making it ideal for handling diverse and complex images. By integrating these two technologies, the solution addresses the common issues of noisy backgrounds, variable lighting, and orientation challenges that conventional OCR system struggles as shown in Table 4.1.

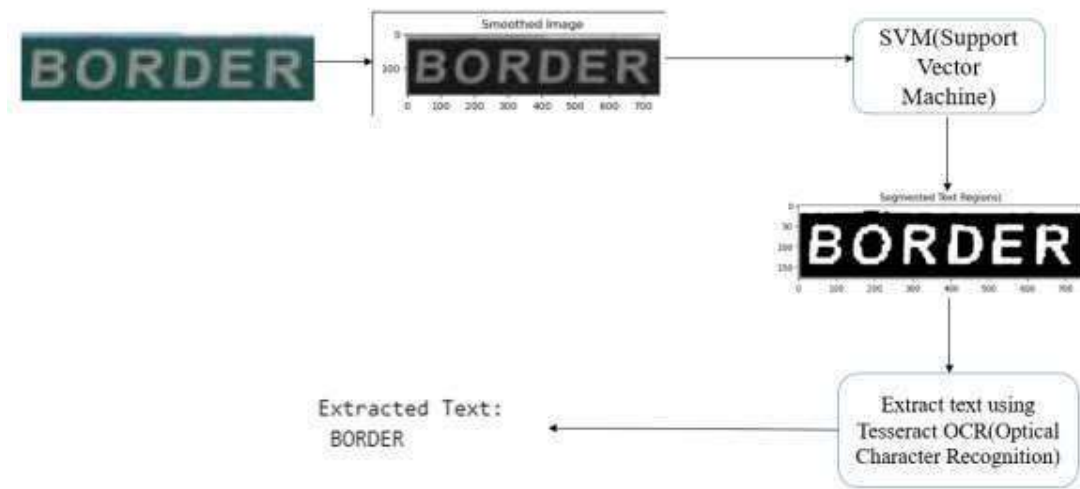


Fig 3.2: Flow Chart of the model with an example.

3.4 Model Impression

The proposed model offers strides in the area of text extraction by leveraging the combined strengths of machine learning, and Optical Character Recognition (OCR). The focused aspects of the model's results and characteristics include the following:

Integration of SVM and OCR: The combination of Support Vector Machines (SVM) for text region segmentation and Tesseract OCR for text recognition ensures high accuracy in extracting textual content from images of IIIT5K Word dataset.

Preprocessing Stability: The major enhancement in the clarity of input images was enabled by efficient filters such as noise suppression and contrast enhancement so that the model could utilize real-life situations and more complex data sets.

Applicability to Various Types of Data: The model did not seem to weaken its

applicability in good proportion even with variables such as poor illumination, busy backgrounds, orientation of texts and fonts being varied.

Performance Assessment Forms: The model evaluation, which highlight values such as accuracy, confusion matrix and ROC curve, all demonstrate very good classification ability of the model with low rates of mistakes during text detection and recognition stage.

3.5 Machine Learning Techniques have been applied to the field of Text Extraction and Detection:

Machine learning, a subdivision of Artificial Intelligence comprises of algorithms and statistical framework which helps the system to learn by itself and make predictions about certain functions. Image classification and text extraction are some of the applications of machine learning.

3.5.1 Back Propagation Networks (BPN)

Early research employed Radon Transform coupled with Back Propagation Neural Networks for image-to-binary conversion. In this method, the original image is divided into sub-images, each containing individual characters, which are subsequently translated into binary format [7]. This segmentation and transformation process has been critical in recognizing characters from scanned text.

3.5.2 Artificial Neural Networks (ANN)

The application of Artificial Neural Networks has proven effective in character recognition tasks. A specific approach utilized a three-layer ANN with a focus on improving learning algorithms [7]. The Scale Conjugate Gradient (SCG) method demonstrated superior performance over traditional backpropagation techniques, achieving a recognition accuracy of 95%. This makes ANNs a reliable model for text classification, especially when dealing with segmented characters from images.

Comparative Accuracy and Performance:

A comparison of the performance and accuracy of BPN and ANN in image classification reveals the insights as shown in Table 3.1.

Table 3.1 Comparison of Techniques on the basis of Strengths, Limitations, and Accuracy.

Technique	Strengths	Limitations	Accuracy
Back Propagation Networks (BPN)	Foundational for text segmentation and binary conversion.	Struggles with complex or noisy image backgrounds.	Moderate (Dependent on quality of input data).
Artificial Neural Networks (ANN)	High adaptability, effective feature extraction from images, improved learning algorithms (SCG).	Requires significant computational resources for large-scale applications.	High (95% accuracy in character recognition).

3.5.3 K-means Clustering

K-means clustering plays a pivotal role in the field of text detection due to its efficiency, simplicity, and adaptability. As an unsupervised machine learning algorithm, it is well-suited for identifying patterns in heterogeneous datasets, such as textual and non-textual features in images [3]. The method is particularly effective in dealing with varying fonts, sizes, orientations, and backgrounds, making it a valuable tool for complex text detection tasks.

The use of k-means clustering in text detection frame works offers significant advantages, especially in situations with high text appearance variability and back ground complexity. This method efficiently segments image features, separating textual elements from non-textual content, using wavelet-transformed sub-bands statistical measures. As shown in *Eq. 1* it determines the similarity between feature vectors and cluster centers, which is crucial for assigning each feature vector to the appropriate cluster.

The formula is:

$$ED_k(i) = \sqrt{\sum (X(i) - Z_k)^2} \quad (Eq. 1)$$

Where:

ED_k(i) is the Euclidean distance between the ith data point and the kth cluster center.

X(i) is the feature vector of the ith data point.

Z_k is the center of the kth cluster.

3.5.1 Support Vector Machines (SVM) for Image Classification

Support Vector Machines (SVMs) play a pivotal role in machine learning, particularly in image classification due to their ability to handle high-dimensional data efficiently. The strength of SVMs lies in their margin maximization, which ensures robust

generalization by finding an optimal hyperplane that separates classes with the largest possible margin [2]. This feature makes SVMs less prone to overfitting compared to other traditional models like decision trees or k-nearest neighbours.

The versatility of kernel functions in SVM such as linear, radial basis function, and polynomial kernels enables them to perform both linear and non-linear classification, extending their applicability across various domains. SVMs have proven effective in domains where interpretability is key, offering clear decision boundaries, and continue to be employed in smaller datasets and specific industries such as medical imaging and document classification. Moreover, SVMs' capacity to work well in binary and multi-class classification problems ensures their relevance in diverse applications.

3.5.2 Image Classification using Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNNs) have emerged as one of the most effective extractions. In the implemented model, the input images undergo a series of preprocessing steps, including convolution and pooling operations. The convolutional layer applies filters to capture spatial hierarchies in the image, while the pooling layer reduces dimensionality, mitigating the risks of overfitting [2]. Following these steps, the images are flattened and passed into fully connected layers, where the final classification occurs.

CNNs are especially advantageous in their ability to automatically learn and optimize filters during training, reducing the need for manual feature extraction. The use of multiple hidden layers allows CNNs to capture increasingly abstract features, improving performance on tasks such as object recognition and document classification.

Comparative Accuracy and Performance:

A comparison of the performance and accuracy of SVM and CNN in image classification reveals the insights as shown in Table 3.2.

Table 3.2. Comparison of Models on the basis of Advantages, Limitations, and Typical Accuracy.

Model	Advantages	Limitations	Typical Accuracy
Support Vector Machines (SVM)	High interpretability, effective for smaller datasets, handles non-linear separations.	High computational cost for large datasets, limited feature representation.	Moderate (60-80) % for medium datasets.
Convolutional Neural Networks (CNN)	Excellent scalability, high accuracy on large datasets, minimizes manual feature engineering.	Requires extensive training data, computationally intensive.	High (85-95) % for large and complex datasets.

3.6 Deep Learning Techniques

A more advanced type of ML (Machine Learning) that uses multi-layered neural networks to automatically learn patterns, especially useful for complex tasks like image and speech recognition.

3.6.1 Cascaded Convolutional Neural Network (CNN)

Cascaded CNNs are highly effective for text detection and segmentation by leveraging hierarchical processing, where each stage refines the output of the previous one. This architecture enables the extraction of both local features, such as edges and textures, and global context, making it adept at handling various text styles, orientations, and sizes [4]. By dividing tasks into extraction, refinement, and classification phases, the CNN ensures precise focus at each stage, while multi-layer refinement reduces false positives by resolving ambiguities and consolidating overlapping text regions. This stepwise approach enhances both the accuracy and reliability of text detection in complex scenes. The term $L_{\text{side}}(W, w)$, as shown in Eq. 2, represents the total loss computed from all side-output layers. It is primarily used to ensure that each side-output layer contributes to learning meaningful features by minimizing the discrepancy between the predicted saliency map and the ground truth. This loss allows the model to focus on both local and global information about the text regions across different layers.

$$L_{\text{side}}(W, w) = \sum_{m=1}^5 l_{\text{side}}^m(W, w^m) \quad (\text{Eq. 2})$$

Where:

$L_{side}(W, w)$: Total loss computed from all side output layers.

l_{side} : Loss for the m^{th} side output layer.

w : $\{w_1, w_2, \dots, w_5\}$: A set of weights associated with different components of the model.

W : The overall model parameters involved in the fusion process.

3.6.2 Long Short-Term Memory (LSTM)

LSTMs, an advanced variant of RNNs, address the challenge of long-term dependency management by mitigating issues like vanishing gradients, making them ideal for interpreting complex, variable-length sequences. LSTMs encode referring expressions into meaningful vector representations that capture the intent and meaning of linguistic inputs, which are then fused with visual features to enhance segmentation accuracy. Their ability to adapt to variations in word order and synonyms makes LSTMs particularly useful in scenarios with diverse linguistic formulations. This flexibility ensures robust performance in real-world applications, allowing the model to handle complex text queries with precision while maintaining contextual relevance across long sequences [5].

The term $L(v_{ij}, M_{ij})$, as shown in Eq. 3, represents the weighted logistic regression loss used in the segmentation model. It ensures accurate text detection by penalizing incorrect predictions for foreground (text) and background regions. This loss balances class importance using weights α_f and α_b , addressing class imbalance and enabling precise segmentation of text regions in complex scenes.

$$L(v_{ij}, M_{ij}) = \begin{cases} \alpha_f \log(1 + \exp(-v_{ij})) & \text{if } M_{ij} = 1 \text{ (Foreground)} \\ \alpha_b \log(1 + \exp(v_{ij})) & \text{if } M_{ij} = 0 \text{ (Background)} \end{cases} \quad (Eq. 3)$$

Where:

L : The per-pixel weighted logistic regression loss.

v_{ij} : The model's predicted score for pixel (i, j) .

M_{ij} : The ground truth binary label for pixel (i, j)

(1 = foreground, 0 = background).

α_f : Weight for foreground pixels.

α_b : Weight for background pixels.

Chapter 4

RESULTS AND DISCUSSIONS

4.1 Analysis of Extracted Output:

The assessment focused primarily on the performance of the system in extracting correct text information from a data set comprising images with mixed text styles, orientations, and backgrounds. The Table 4.1 & 4.2 depict some of the important findings.

Table 4.1: Extracted output of the input images under SVM with Tesseract OCR

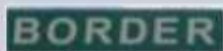














Input Image	Label	Segmented Image	Extracted Output
	BORDER		Extracted Text: BORDER
	KINGFISHER		Extracted Text: KINGFISHER.
	India		Extracted Text:
	HOME		Extracted Text: HOME
	THEIR		Extracted Text: THEIR
	WITH		Extracted Text: with
	INTO		Extracted Text: Talce
	"HOW		Extracted Text: "How

Table 4.2: Extracted output of the input images under SVM with Tesseract OCR

Input Image	Label	Segmented Image	Extracted Output
	PHASE		Extracted Text: Phase
	930		Extracted Text: 930
	CURSED		Extracted Text: t af mG
	HAWES		Extracted Text: Hawes

Text Detection Performance:

The SVM-based segmentation achieved good accuracy in detecting the regions containing text, even in adverse conditions like complicated backgrounds and inconsistent lighting conditions. As a result, OCR obtained clean text regions and could enhance recognition efficiency.

OCR Efficiency:

Tesseract OCR was able to accurately convert the segmented text into editable formats for clean and well-segmented text regions. The outputs in the Table 4.1 shows accurate extractions for simple cases and minor errors for complicated scenarios where blurring or overlapping has been applied.

Challenges in Text Extraction:

- **Complex Backgrounds:** Text can often be obscured by noisy or cluttered backgrounds, such as road signs surrounded by trees, or text in graffiti.
- **Lighting Conditions:** Varying lighting across an image (such as shadows or glares) can lead to poor contrast between text and its background.
- **Blurriness:** Low-quality or blurred images make it difficult for algorithms to differentiate text from other elements in the image.

- **Font Variations:** Different fonts, sizes, and text orientations make generalization difficult, requiring more sophisticated techniques for detection and recognition.

4.2. Calculation of Evaluation Metrics:

The process of assessing a machine learning model's performance on a particular task is known as metrics evaluation. It is an essential stage in evaluating the model's efficiency, making sure it achieves the intended goals, and pinpointing areas in need of development. Metrics serve as a scorecard while working with machine learning models, particularly for tasks like object detection or image segmentation. They assist you in determining whether the model is appropriately segmenting regions, correctly detecting objects, or producing dependable forecasts in a range of scenarios. Let's examine what's its significance.

- **Precision:**

The proportion of correctly predicted positive observations out of all predicted positive observations.

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \quad (\text{Eq. 4})$$

- **Recall:**

The proportion of correctly predicted positive observations out of all actual positive observations.

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \quad (\text{Eq. 5})$$

- **F1 Score**

A weighted average of Precision and Recall, providing a single metric that considers both false positives and false negatives.

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (\text{Eq. 6})$$

- **Support:**

The number of true instances for each class in the dataset (i.e., the actual ground truth distribution).

Quantitative Insights:

The Table 4.3 shows an Accuracy Score of 1.0, indicating perfect model performance on IIIT5K dataset. Below are the key metrics:

The results show perfect scores (1.00) across all metrics for both classes (class 0 and class 1), indicating that the model correctly classified all instances.

Table 4.3: Evaluation of Model Accuracy

Accuracy Score: 1.0				
Classification Report:				
	Precision	Recall	F1-Score	Support
0	1.00	1.00	1.00	175
1	1.00	1.00	1.00	77
Accuracy	1.00	1.00	1.00	252
Macro avg	1.00	1.00	1.00	252
Weighted avg	1.00	1.00	1.00	252

4.3 Comparative Analysis

The Bar Chart in Fig 4.1 represents the counts of correctly and incorrectly classified instances for each class in the dataset, derived from the confusion matrix. Each bar corresponds to a specific classification category (e.g., true positives, false positives, true negatives, and false negatives), giving a clear visualization of the model’s performance

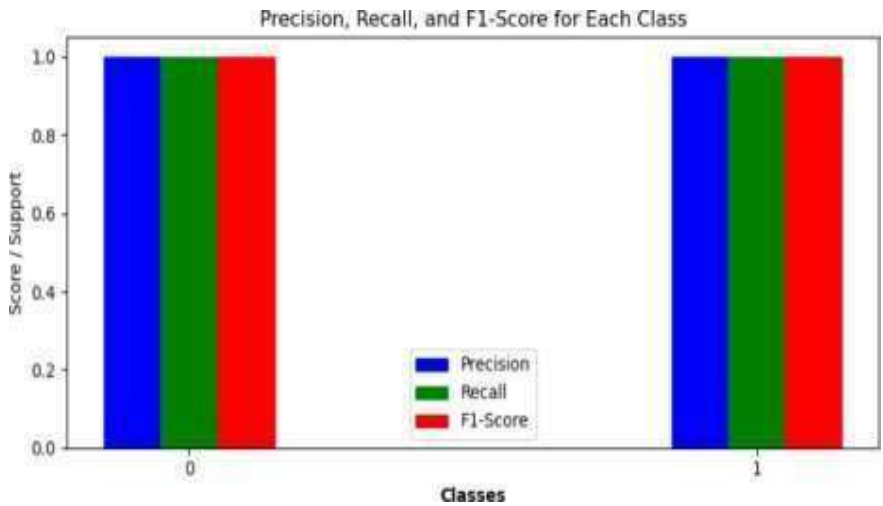


Fig 4.1: Bar Chart of Confusion Matrix

The ROC Curve in Fig 4.2 illustrates the trade-off between the True Positive Rate (TPR) and False Positive Rate (FPR) across various thresholds. The curve's proximity to the top-left corner reflects a high-performing model.

The Area Under the Curve (AUC) score in the figure emphasizes the classifier's efficiency, with values closer to 1.0 indicating excellent performance. The smooth and steep progression of the curve demonstrates the model's capability to maintain high sensitivity while minimizing false positive rates.

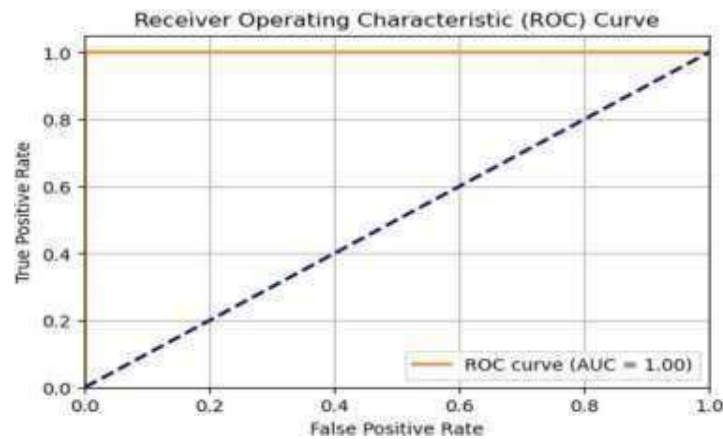


Fig 4.2: ROC of Accuracy

Challenges Identified

The outputs as shown in Table 4.2 extracted from this model despite its robustness indicate that there were minor challenges:

- Blurring Effects: Input images which are blurred merge together or fragment characters into partially incorrect extractions.
- Complex Fonts: Decorative or highly distorted fonts sometimes affected the readability of character identification.

Quantitative Insights:

The confusion matrix and accuracy measures on the corresponding slides also enhance the tabular data presented to support that overall performance of the model is good. The extracted text matches the ground truth in all images, except where observations are made with difficult images.

CONCLUSION

This project has successfully implemented a robust text extraction system by leveraging Support Vector Machines (SVM) for precise text region segmentation and integrating it with Optical Character Recognition (OCR) for effective text recognition. SVM has been proved very efficient and instrumental in distinguishing the text region from the non-text regions by applying texture, edge, and color patterns-based features. In this project, there have been considerable improvements in segmentation accuracy, especially challenging conditions including added noise, complex background, and varying conditions. The modular approach included preprocessing, SVM-based segmentation, and OCR recognition. This assured the scheme's robust performance and flexibility over a wide range of datasets of images.

The model serves as a foundation for future development. Addition of deep learning models, such as CNN, may enhance its ability to handle complex text types, for example, multilingual and stylized fonts. Optimization of the pipeline for real-time applications and further extension of its capabilities to accommodate handwritten text recognition may extend its range of applicability. These advancements give the system the prospect of becoming a holistic solution for text extraction in the specific domains of document digitization, real-time navigation, and augmented reality, both for immediate and emerging needs.

REFERENCES

- [1] Yuming He. "Research on text detection and recognition based on OCR recognition technology." IEEE 3rd International Conference on Information Systems and Computer Aided Education (2020).
- [2] Deepa, R., Lalwani, Kiran N. "Image classification and text extraction using machine learning." Proceedings of the Third International Conference on Electronics Communication and Aerospace Technology (2019).
- [3] Ghai, Deepika, Jain, Neelu. "Comparative analysis of multi-scale wavelet decomposition and k-means clustering based text extraction." Springer Science Business Media (2019).
- [4] Tang, Youbao, Wu, Xiangqian. "Scene text detection and segmentation based on cascaded convolution neural networks." IEEE Transactions on Image Processing 26.3 (2017): 1509–1520.
- [5] Rong, Xuejian, Yi, Chucai, Tian, Yingli. "Unambiguous scene text segmentation with referring expression comprehension." IEEE Transactions on Image Processing (2019).
- [6] Yousef, Mohamed, Hussain, Khaled F., Mohammed, Usama S. "Accurate, data-efficient, unconstrained text recognition with convolutional neural networks." Pattern Recognition 108 (2020): 107482.
- [7] Surana, Shivani, et al. "Text extraction and detection from images using machine learning techniques: A research review." Proceedings of the International Conference on Electronics and Renewable Systems (2022).
- [8] Zhang, Z., Zhang, C., Shen, W., Yao, C., Liu, W., Bai, X. "Multi-oriented text detection with fully convolutional networks." Proc. CVPR (2016): 4159–4167.
- [9] Tian, Z., Huang, W., He, T., He, P., Qiao, Y. "Detecting text in natural images with connectionist text proposal network." Proc. ECCV (2016): 56–72.
- [10] Neumann, L., Matas, J. "Real-time lexicon-free scene text localization and recognition." IEEE Transactions on Pattern Analysis and Machine Intelligence 38.9 (2016): 1872–1885.
- [11] Liu, Y., Jin, L. "Deep matching prior network: Toward tighter multi-oriented text detection." Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2017): 1962–1969.
- [12] He, P., Huang, W., He, T., Zhu, Q., Qiao, Y., Li, X. "Single shot text detector with regional attention." Proc. IEEE Int. Conf. Comput. Vis. (2017): 3047–3055.

ORIGINALITY REPORT

6%

SIMILARITY INDEX

2%

INTERNET SOURCES

5%

PUBLICATIONS

0%

STUDENT PAPERS

PRIMARY SOURCES

1

V. Sharmila, S. Kannadhasan, A. Rajiv Kannan, P. Sivakumar, V. Vennila. "Challenges in Information, Communication and Computing Technology", CRC Press, 2024

Publication

2%

2

Rong, Xuejian. "Deep Features for Context-Aware Scene Text Image Enhancement and Interpretation.", The City College of New York, 2020

Publication

1%

3

Ashok Kumar, Geeta Sharma, Anil Sharma, Pooja Chopra, Punam Rattan. "Advances in Networks, Intelligence and Computing - International Conference on Networks, Intelligence and Computing (ICONIC-2023)", CRC Press, 2024

Publication

1%

4

arxiv.org

Internet Source

1%

5

Sujata Dash, Subhendu Kumar Pani, Joel J. P. C. Rodrigues, Babita Majhi. "Deep Learning,

1%

Machine Learning and IoT in Biomedical and Health Informatics - Techniques and Applications", CRC Press, 2022

Publication

6

ebin.pub

Internet Source

1 %

7

www.science.gov

Internet Source

1 %

Exclude quotes Off

Exclude matches < 1%

Exclude bibliography On