##### A Project report on

**Object Detection and Speech Recognition Using Machine Learning**

###### A Dissertation submitted to JNTU Hyderabad in partial fulfillment of the academic requirements for the award of the degree.

**Bachelor of Technology**

**in**

**Computer Science and Engineering**

Submitted by

A.Rohan Reddy

(19H51A0563)

B.Pradeep

(20H51A0584)

Riyaz Ahmed

(20H51A05L6)

Under the esteemed guidance of

Mr.V. Narasimha

(Assistant Professor)



**Department of Computer Science and Engineering**

**CMR COLLEGE OF ENGINEERING & TECHNOLOGY**

(UGC Autonomous)

\*Approved by AICTE \*Affiliated to JNTUH \*NAAC Accredited with A+ Grade

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD - 501401.

#### 2020- 2024

**CMR COLLEGE OF ENGINEERING & TECHNOLOGY**

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD – 501401

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**



#### CERTIFICATE

This is to certify that the Major Project Phase I report entitled **"Object Detection and Speech Recognition Using Machine Learning"** being submitted by A. Rohan Reddy (19H51A0563), B. Pradeep (20H51A0584), Riyaz Ahmed(20H51A0L6) in partial fulfillment for the award of **Bachelor of Technology in Computer Science and Engineering** is a record of bonafide work carried out his/her under my guidance and supervision.

###### The results embodies in this project report have not been submitted to any other University or Institute for the award of any Degree.

**Mr.V.Narasimha Dr. Siva Skandha Sanagala**

**Assistant Professor Associate Professor and HOD**

**Dept. of CSE Dept. of CSE**

#### ACKNOWLEDGEMENT

With great pleasure we want to take this opportunity to express my heartfelt gratitude to all the people who helped in making this project work a grand success.

We are grateful to **Mr.V.Narasimha, Assistant Professor** , Department of Computer Science and Engineering for his valuable technical suggestions and guidance during the execution of this project work.

We would like to thank **Dr. Siva Skandha Sanagala,** Head of the Department of Computer Science and Engineering, CMR College of Engineering and Technology, who is the major driving forces to complete my project work successfully.

We are very grateful to **Dr. Vijaya Kumar Koppula**, Dean-Academics, CMR College of Engineering and Technology, for his constant support and motivation in carrying out the project work successfully.

We are highly indebted to **Major Dr. V A Narayana,** Principal, CMR College of Engineering and Technology, for giving permission to carry out this project in a successful and fruitful way.

We would like to thank the **Teaching & Non- teaching** staff of Department of Computer Science and Engineering for their co-operation

We express our sincere thanks to **Shri. Ch. Gopal Reddy**, Secretary, CMR Group of Institutions, for his continuous care.

Finally, We extend thanks to our parents who stood behind us at different stages of this Project. We sincerely acknowledge and thank all those who gave support directly and indirectly in completion of this project work.

A.Rohan Reddy 19H51A0563

B.Pradeep 20H51A0584

Riyaz Ahmed 20H51A05L6

**TABLE OF CONTENTS**

**CHAPTER**

**NO. TITLE PAGE NO.**

ABSTRACT iv

**1** **INTRODUCTION** 1

1.1 Problem Statement 2

1.2 Research Objective 2

1.3 Project Scope and Limitations 3

**2** **BACKGROUND WORK** 4

2.1. Histogram of Oriented Gradients (HOG) 4

2.1.1. Introduction 4

2.1.2. Merits, Demerits and Challenges 6

2.1.3. Implementation of Histogram of Oriented Gradients (HOG) 9

2.2. R-CNN 12

2.2.1. Introduction 12

2.2.2. Merits, Demerits and Challenges 14

2.2.3. Implementation of R-CNN 16

2.3. Single shot detector (SSD) 19

2.3.1. Introduction 19

2.3.2. Merits, Demerits and Challenges 22

2.3.3. Implementation of Single shot detector (SSD) 24

**3 RESULTS AND DISCUSSION 42** 4.1. Performance metrics 42

4 **CONCLUSION** 50

5.1 Conclusion 50

**REFERENCES** **52**

CMRCET B. Tech (CSE) Page No 1

Object Detection and Speech Recognition Using Machine Learning

# **ABSTRACT**

In recent years, machine learning techniques have revolutionized various fields of technology, paving the way for innovative applications and solutions. This abstract delves into two prominent applications of machine learning:

Object Detection, a subfield of computer vision, enables computers to identify and locate objects within images or video frames. Leveraging deep learning models such as Convolutional Neural Networks (CNNs), object detection algorithms have achieved unprecedented accuracy. This capability has found numerous practical applications, including autonomous vehicles, surveillance systems, and augmented reality.

.

CMRCET B. Tech (CSE) Page No 3

Object Detection and Speech Recognition Using Machine Learning

# **CHAPTER 1**

**INTRODUCTION**

CMRCET B. Tech (CSE) Page No 1

Object Detection and Speech Recognition Using Machine Learning

**CHAPTER 1**

**INTRODUCTION**

**1.1.Problem Statement**

Object Detection is a fundamental subfield of computer vision, focusing on the task of locating and identifying objects within images or video frames. This technology has undergone a remarkable transformation, largely owing to the rise of deep learning, particularly Convolutional Neural Networks (CNNs). These networks have demonstrated exceptional capabilities in recognizing objects with precision and speed, making object detection a critical component in applications such as self-driving cars, security surveillance systems, and augmented reality experiences.

Speech Recognition and vice versa, on the other hand, is the art of converting spoken language into text or executable commands. In recent years, machine learning techniques, especially recurrent neural networks (RNNs) and advanced Transformer-based models, have fueled substantial progress in this field. Speech recognition systems have evolved to a point where they can understand and transcribe spoken words with high accuracy. This has led to the proliferation of virtual assistants like Siri and Alexa, as well as the development of tools for transcription services, voice assistants, and accessibility applications, making it easier for people to communicate with computers and other devices through natural language.

.

**1.2.Research Objective**

**Object Detection:**

1. **Improving Accuracy and Speed**: Enhance the accuracy and speed of object detection algorithms, especially in real-time or resource-constrained scenarios.
2. **Handling Challenging Conditions**: Develop methods to make object detection more robust under challenging conditions such as low light, occlusion, or adverse weather.

CMRCET B. Tech (CSE) Page No 2

Object Detection and Speech Recognition Using Machine Learning

**1.3. Project Scope and Limitations**

**Project Scope:**

1. Autonomous Vehicles: Object recognition plays a crucial role in the development of autonomous vehicles. It helps in identifying and classifying objects on the road, such as pedestrians, other vehicles, traffic signs, and obstacles. The future of transportation heavily relies on accurate and efficient object recognition systems to ensure safety.

2.Augmented Reality (AR) and Virtual Reality (VR): Object recognition is fundamental in AR and VR technologies. These technologies rely on the ability to recognize and interact with real-world objects or environments. The future holds potential for more immersive and interactive experiences in gaming, education, training, and various industries.

3.Healthcare: Object recognition can be utilized in medical imaging for the identification of anomalies, tumors, or specific organs. It aids in diagnosis and treatment planning. Additionally, it can be used for tracking and identifying medical tools and devices in a clinical setting. .

**Limitations:**

1. **Detection of Small Objects**: Detecting an small object within an image is challenging, especially when their features are not clear or they’re surrounded by other larger objects
2. **Variability in Object Appearance:** Objects can appear differently due to variations in scale, rotation,lighting,occlusion, and deformation, making it challenging for detectors to identify them accurately.
3. **Complex Backgrounds and Clutter:** Busy or Cluttered backgrounds can confuse object detection models, leading to false positives or missed detections.
4. **Limited Object Categories:** Object detection models might be limited in their ability to rare or new object categories not present in their training data

CMRCET B. Tech (CSE) Page No 3

Object Detection and Speech Recognition Using Machine Learning

**CHAPTER 2**

**BACKGROUND WORK**

CMRCET B. Tech (CSE) Page No 4

Object Detection and Speech Recognition Using Machine Learning

**CHAPTER 2**

**BACKGROUND WORK**

**2.1 Histogram of Oriented Gradients (HOG)**

**2.1.1 Introduction**

HOG is a feature descriptor used in object detection within images. Proposed by Navneet Dalal and Bill Triggs in 2005.This method gained popularity ,especially in pedestrian detection and other object detection applications

HOG operates by dividing an image into small regions, computing the gradient orientation

and magnitude within in this region, and then creating histograms of gradients for these regions.

These histograms represent the distribution of gradient orientations, which help characterize the shape and appearance of object

**2.1.2 Merits, Demerits and Challenges**

**Merits:**

**1.Simple and Efficient Feature Descriptor:** HOG provides a straightforward yet effective method for describing local object structure and appearance in images. It’s computationally

efficient, making it feasible for real-time applications.

**2.Robust to illumination changes**: HOG is relatively robust to changes in lighting and contrast

variations, as it focuses on local gradient orientations rather than absolute pixel values.

**3.Effective for pedestrian detection:** Particularly successful in pedestrian detection tasks, HOG features have been widely used in scenarios where the shape and structure of the object are essential, such as identifying humans in images

CMRCET B. Tech (CSE) Page No 5

Object Detection and Speech Recognition Using Machine Learning

**Demerits:**

**1.Sensitivity to Variations in Scale and Rotation:** HOG is sensitive to changes in object scale and rotation, making it less effective in scenarios where objects appear at different scales or orientations.

**2.Limited Adaptability to Occlusions:** When objects are occluded or partially hidden, HOG may struggle to accurately detect and localize the complete object due to its reliance on local gradient-based features.

**Challenges:**

**1.Variations in Scale and Rotation**: HOG is sensitive to changes in the scale and rotation of objects, leading to reduced accuracy when objects appear in different sizes or orientations.

**2.Occlusion Handling**: Detecting partially occluded objects is a challenge for HOG. When objects are partially obscured or overlapped by other objects, it can struggle to accurately detect and localize the entire object.

**2.1.3 Implementation of HOG:**

Data Collection and Preprocessing:

Gather a dataset containing images with the objects you want to detect. Ensure the images are annotated with bounding boxes around the objects of interest. Preprocess the images (resize, normalize, etc.) for consistency.

Image Gradient Calculation:

Compute the gradient magnitude and direction for each pixel in the image. This is often done by using techniques like the Sobel operator to calculate the gradient in both the x and y directions.

Gradient Histogram Computation:

Divide the image into small regions called cells. For each pixel in a cell, calculate the gradient orientation and magnitude. Accumulate these pixel gradients into orientation histograms for the cell.

CMRCET B. Tech (CSE) Page No 6

Object Detection and Speech Recognition Using Machine Learning

Block Normalization:

Group cells into larger blocks. Normalize the histograms within each block to handle changes

in lighting and contrast. Normalization could involve methods such as L2 normalization.

Descriptor Blocks:

Concatenate the normalized histograms from all the blocks to form the final descriptor for the image. The descriptor contains information about the object's gradients within the image.

Training a Classifier:

Use the generated HOG descriptors along with their corresponding annotated bounding boxes to train a machine learning model (commonly a support vector machine, SVM) or other classifiers. This step helps the model learn to distinguish between positive (object present) and negative (background) samples.

Sliding Window and Detection:

Slide a window across the image at different scales and positions. Extract HOG descriptors from each window and feed them into the trained classifier. If a region is classified as containing an object, mark it as a detection.

Non-maximum Suppression:

Remove duplicate or overlapping detections by applying non-maximum suppression. This step helps to retain the most confident and accurate detections by suppressing weaker, overlapping ones.

Testing and Evaluation:

Test the object detector on new, unseen images. Evaluate its performance using metrics like precision, recall, and F1-score. Adjust parameters or optimize the model if needed.

CMRCET B. Tech (CSE) Page No 7

Object Detection and Speech Recognition Using Machine Learning

**2.2 R-CNN**

**2.2.1 Introduction**

Region-based Convolutional Neural Network (R-CNN) is a significant advancement in the field of computer vision, particularly in the area of object detection and semantic segmentation. Developed by Ross Girshick and his team in their groundbreaking paper published in 2014, R-CNN is a framework designed to accurately identify objects within images and precisely localize them. It has since become a foundational technique for a wide range of applications, including autonomous driving, image recognition, and more.

R-CNN addresses the complex task of object detection as a multi-step process:

Region Proposals: The first step involves generating region proposals. These are potential bounding boxes that might contain objects. Various methods, such as selective search, are used to create these proposals, which serve as candidates for objects within the image.

Feature Extraction: For each region proposal, a pretrained Convolutional Neural Network (CNN) is utilized to extract deep features. The CNN is typically pretrained on a large dataset for general object recognition tasks and then fine-tuned specifically for the object detection task at hand.

Classification and Localization: The extracted features are used for object classification and localization. A set of class-specific Support Vector Machines (SVMs) is employed for object classification, determining which object category each region corresponds to. Additionally, bounding box regression is applied to refine the location of the objects within the proposed regions. This step ensures a precise localization of the objects.

CMRCET B. Tech (CSE) Page No 7

Object Detection and Speech Recognition Using Machine Learning

**2.2.2 Merits, Demerits and Challenges**

**Merits:**

1. High Accuracy: One of the standout merits of R-CNN is its exceptional accuracy in object detection. It significantly outperformed earlier object detection methods, setting new standards in accuracy and reliability.
2. Versatility: R-CNN is a versatile framework that can be applied to a wide variety of object detection tasks. It is not limited to specific object categories and performs well in generic object detection.
3. Precise Localization: The localization accuracy of R-CNN is a major advantage. The bounding box regression step plays a crucial role in ensuring precise object localization.

**Demerits:**

1. Computational Overhead: R-CNN is computationally intensive due to its step-by-step approach. Extracting features for each region proposal individually can be time-consuming and resource-intensive. This computational overhead limits its suitability for real-time applications and environments where low latency is critical.

**Challenges:**

1. Complex Training Process: Training an R-CNN model is a complex and resource-intensive process. It involves multiple stages, including fine-tuning a pretrained CNN, training class-specific SVM classifiers, and training bounding box regression models. This complexity can be challenging for practitioners and researchers, requiring access to substantial computational resources.
2. Region Proposal Limitation: R-CNN relies on external region proposal methods, such as selective search. These methods, while effective, introduce a bottleneck in the pipeline. Selective search, for example, can be slow and may not always provide the best region proposals. Overcoming this limitation and improving the efficiency of region proposal generation is an ongoing challenge in the development of object detection techniques.

CMRCET B. Tech (CSE) Page No 8

Object Detection and Speech Recognition Using Machine Learning

**2.2.3 Implementation of R-CNN**

Implementing R-CNN involves a series of steps:

Data Preparation: Start by collecting and preparing the dataset, which includes images and annotations that specify the objects present in the images.

Select Region Proposal Method: Choose a region proposal method, such as selective search, to generate candidate regions for object detection. These proposals will serve as the initial candidates for objects in the images.

Pretrained CNN Fine-Tuning: Select a pretrained Convolutional Neural Network (CNN) model, such as AlexNet or VGGNet, and fine-tune it using the collected dataset. This fine-tuning adapts the network to the specific object detection task, allowing it to learn object-specific features.

Train Object Classifiers: Train class-specific SVM classifiers to identify objects within the region proposals. These classifiers will determine which object category each proposed region corresponds to.

Train Bounding Box Regressors: Train bounding box regressors to refine the localization of objects. This step helps improve the accuracy of object bounding box predictions.

Inference: During the inference stage, apply the complete R-CNN pipeline to new images. This includes region proposal generation, feature extraction, object classification, and bounding box regression to detect and locate objects within the images.

While R-CNN marked a significant advancement in object detection using deep learning, it has evolved into more efficient and faster models. Notable successors to R-CNN include Fast R-CNN, Faster R-CNN, and YOLO (You Only Look Once).

CMRCET B. Tech (CSE) Page No 9

Object Detection and Speech Recognition Using Machine Learning

**2.3 Single shot detector (SSD)**

**2.3.1 Introduction**

Single Shot Detector (SSD) is a groundbreaking object detection model in computer vision, designed to identify and localize objects in images with high accuracy and real-time performance. SSD was introduced by Wei Liu, et al. in 2016. It stands out for its ability to simultaneously predict multiple object categories and their corresponding bounding boxes in a single forward pass of a neural network.

SSD addresses the limitations of previous object detection models by combining the strengths of deep convolutional neural networks (CNNs) with efficient techniques for predicting object classes and locations. It achieves impressive results in terms of accuracy and speed, making it a popular choice for various applications, including autonomous vehicles, surveillance, and image analysis.

**2.3.2** **Merits, Demerits and Challenges**

**Merits:**

1. High Accuracy: SSD offers high levels of accuracy in object detection. It outperforms many earlier models in terms of precision and recall, which are critical for object detection tasks.
2. Real-Time Performance: One of the primary merits of SSD is its ability to provide real-time or near real-time object detection. This is crucial for applications that require fast processing, such as autonomous vehicles and video surveillance.
3. Multi-Scale Detection: SSD excels at detecting objects at various scales within an image. It can identify both small and large objects in a single pass, making it suitable for diverse scenarios.

CMRCET B. Tech (CSE) Page No 9

Object Detection and Speech Recognition Using Machine Learning

1. Localization Precision: The model provides precise object localization, which is vital for applications that require accurate bounding box predictions.

**Demerits:**

Training Complexity: Training an SSD model can be complex and resource-intensive, especially when dealing with large-scale datasets and a wide range of object categories. Fine-tuning a pretrained model for specific tasks can require substantial computational resources. **Challenges:**

1. Trade-off between Speed and Accuracy: While SSD is known for its real-time performance, there is often a trade-off between speed and accuracy. Achieving the highest accuracy may require sacrificing some speed, which can be a challenge in applications that demand both.
2. Handling Occlusions and Crowded Scenes: SSD, like many object detection models, may struggle with accurately detecting objects in crowded scenes or when objects are partially occluded. This remains a challenge in the field of object detection.

**2.2.3 Implentation of Single shot detector (SSD)**

Implementing SSD involves several key steps:

Data Preparation: Begin by collecting and preparing the dataset, which includes images and annotations specifying object categories and bounding box locations.

Model Selection: Choose a suitable SSD architecture variant and a pretrained CNN model, such as VGGNet or ResNet, as the base network. Depending on your specific task and computational resources, you can select an appropriate SSD variant.

Training: Fine-tune the chosen model using the collected dataset. Training involves optimizing the model's weights for object detection. This step can be time-consuming and computationally demanding.

CMRCET B. Tech (CSE) Page No 11

Object Detection and Speech Recognition Using Machine Learning

Inference: During the inference phase, apply the trained SSD model to new images. The model predicts object categories and bounding boxes in real-time or near real-time, depending on the chosen model variant.

Post-processing: After inference, you can apply post-processing techniques to filter and refine the detected objects and their bounding boxes. This step helps improve the quality of the results.

CMRCET B. Tech (CSE) Page No 11

Object Detection and Speech Recognition Using Machine Learning

**CHAPTER 3**

**RESULTS AND DISCUSSION**

CMRCET B. Tech (CSE) Page No 12

Object Detection and Speech Recognition Using Machine Learning

**CHAPTER 3**

**RESULTS AND DISCUSSION**

The Object Detection and Speech Recognition Using Machine Learning is currently in the research phase, and our preliminary findings and discussions revolve around the examination of existing solutions addressing the same problem statement. Our research efforts have been dedicated to understanding the landscape of solutions available for identifying an individual's mother tongue from their English speech.

A central focus of our research has been the analysis of existing solutions and methodologies employed in similar domains. These solutions span a spectrum of techniques, ranging from traditional machine learning approaches to state-of-the-art deep learning models. This comprehensive analysis has provided valuable insights into the strengths and weaknesses of these existing solutions.

One key observation that has emerged from our research is the delicate balance between model complexity and accuracy. Many existing solutions leverage intricate deep learning architectures to achieve high levels of accuracy. However, this high accuracy often comes at the cost of increased model complexity, making deployment in real-world applications a resource-intensive task.

Another notable challenge evident in the landscape of existing solutions is their capacity to perform accurately in multilingual environments. Given the diverse linguistic backgrounds and varying levels of English fluency among speakers, accurately identifying the speaker's mother tongue becomes a multifaceted challenge.

Our discussions have also encompassed the practicality and ease of integration of existing solutions. Some solutions are tailored for specific applications and may lack versatility, while others offer flexibility for integration into a range of contexts.

CMRCET B. Tech (CSE) Page No 13

Object Detection and Speech Recognition Using Machine Learning

CHAPTER 4

**CONCLUSION**

CMRCET B. Tech (CSE) Page No 14

Object Detection and Speech Recognition Using Machine Learning

**CHAPTER 4**

**CONCLUSION**

In conclusion, the Object Detection and Speech Recognition Using Machine Learning, currently in the research phase, represents a pioneering effort to tackle the intricate task of identifying an individual's mother tongue from their English speech patterns. Our ongoing research, findings, and discussions have shed light on the model's significance and potential applications. Through a comprehensive analysis of existing solutions, we have observed the delicate balance between model complexity and accuracy, recognizing the challenges associated with multilingual environments. The model has shown initial promise with commendable accuracy, albeit with challenges when dealing with highly fluent English speakers. As we look forward, our commitment to enhancing accuracy, adaptability, and practicality remains steadfast. The model's impact extends to language learning tools, cross-cultural communication, and speech analysis, offering the potential to bridge linguistic gaps and provide valuable insights into the subtleties of the human voice. It is not merely a research project; it represents a path toward a more inclusive and interconnected global society. As we refine and expand its capabilities, we envision a future where the Object Detection and Speech Recognition Using Machine Learning plays a pivotal role in language education, cross-cultural understanding, and unraveling the intricate tapestry of accents and languages that define our world.

CMRCET B. Tech (CSE) Page No 15

Object Detection and Speech Recognition Using Machine Learning

**REFERENCES**

CMRCET B. Tech (CSE) Page No 16

Object Detection and Speech Recognition Using Machine Learning

## REFERENCES

1. <https://scholarworks.calstate.edu/downloads/qz20sv442>
2. <https://medium.com/analytics-vidhya/using-machine-learning-to-identify-accents-in-spectrograms-of-speech-5db91c191b6b>
3. <https://www.mdpi.com/2227-7390/10/16/2913>
4. <https://www.audeering.com/research/opensmile/>
5. <https://www.ispeech.org/>
6. <https://cloud.google.com/speech-to-text>

## 

CMRCET B. Tech (CSE) Page No 17