RSET
RAJAGIRI SCHOOL OF
ENGINEERING & TECHNOLOGY
(AUTONOMOUS)

Project Phase II Report On

MoRedact: A Motion-Picture Censorship Application

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

Bachelor of Technology

in

Computer Science and Engineering

By

Ameen Mohammed (U2003033)
Adithya M (U2003214)
Ajith Bobby (U2003014)
Ankit John Abraham (U2003037)

Under the guidance of

Mr. Harikrishnan M.

Department of Computer Science and Engineering
Rajagiri School of Engineering & Technology (Autonomous)
(Parent University: APJ Abdul Kalam Technological University)
Rajagiri Valley, Kakkanad, Kochi, 682039
April 2024

CERTIFICATE

This is to certify that the project report entitled ***"MoRedact: A Motion-Picture Censorship Application"*** *is a bonafide record of the work done by* ***Ameen Mohammed (U2003033), Adithya M (U2003214), Ajith Bobby (U2003014), Ankit John Abraham (U2003037)****, submitted to the Rajagiri School of Engineering & Technology (RSET) (Autonomous) in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology (B. Tech.) in Computer Science and Engineering during the academic year 2023-2024.*

Mr. Harikrishnan M
Project Guide
Asst. Professor
Dept. of CSE
RSET

Dr. Sminu Izudheen
Project Coordinator
Professor
Dept. of CSE
RSET

Dr. Preetha K G
Head of the Department
Professor
Dept. of CSE
RSET

ACKNOWLEDGMENT

We wish to express our sincere gratitude towards **Dr. Sreejith P S**, Principal of RSET, and **Dr. Preetha K G**, Head of the Department of "Computer Science and Engineering" for providing us with the opportunity to undertake our project, "MoRedact: A Motion-Picture Censorship Application".

We are highly indebted to our project coordinators, **Dr. Sminu Izudheen**, Professor, Dept. of CSE, and **Dr. Renu Mary Daniel**, Asst. Professor, Dept. of CSE, for their valuable support.

It is indeed our pleasure and a moment of satisfaction for us to express our sincere gratitude to our project guide **Mr. Harikrishnan M** for his patience and all the priceless advice and wisdom he has shared with us.

Last but not the least, we would like to express our sincere gratitude towards all other teachers and friends for their continuous support and constructive ideas.

Ameen Mohammed

Adithya M

Ajith Bobby

Ankit John Abraham

Abstract

The film industry is always changing, and the rapid growth of digital media and streaming platforms has made sophisticated content regulation procedures necessary. Having realised this necessity, our project "Automation in Movie Censorship" aims to expand the definition of content supervision by incorporating state-of-the-art modules that will transform the way we oversee and manage cinematic entertainment.

The Human Action Recognition Module is in the front of our technology arsenal. This module uses a convolutional neural network (CNN) architecture with two streams that are split into temporal and spatial networks on purpose. The temporal network examines video sequences to find patterns over time, while the spatial network painstakingly pulls characteristics from each frames. The combination of these attributes enables the accurate categorization of human behaviours, providing a sophisticated method of content evaluation that surpasses traditional techniques.

To enhance the overall experience, our proposal includes a novel Inappropriate Content in Image Detection Module that utilises a real-time censorship algorithm based on YOLO. This technique uses a pipeline-based design, which parallelizes operations with subprocesses to increase efficiency. By using this advanced method, the system quickly determines if a picture or a frame from a movie is appropriate or inappropriate, guaranteeing a proactive and efficient content moderation procedure.

Our project presents the Inappropriate Audio Censorship Module, which delves deeper into the auditory domain. The module takes advantage of the Whisper ASR for identification of speech in the video input. The audio is separated, speech is identified, along with the timestamps for the occurrences of each word found. Further processing of the audio extracted is done using the Fast Forward Moving Picture Experts Group

(FFMPEG), which enables for the muting/censoring of the words found in the audio that are deemed explicit. Additionally, creation of the transcripts from the audio is used as the primary input for the genre classification module.

We further extend our reach by adding the Genre Classification Module, which employs an NLP based approach to extract textual features from the video transcript using TF-IDF and vector embeddings obtained through Word2Vec. The resultant vector format datat is passed to the proposed model involving a Multinomial Naive Bayes approach, to predict the multi-label classified genres.

As we set out on this technological journey, our endeavour represents a radical change in how we view film censorship. Our goal is to create a cinematic world that is not only entertaining but also considerate of the varying sensitivities of its audience by embracing the potential of artificial intelligence and sophisticated algorithms. We go into great detail about each module in the pages that follow, emphasising the creative approaches used and how they can affect future content regulation in the film industry.

Contents

List of Abbreviations

CNN - Convolutional Neural Network

RNN - Recurrent Neural Networks

ASR - Automatic Speech Recognition

ResNet - Residual Network

MFCC - Mel Frequency Cepstral Coefficients

GRU - Gated Recurrent Unit

SVM - Support Vector Machine

SSD - Single Shot Detector

LBP - Local Binary Pattern

C3D - Convolutional 3D Neural Network

LRCN - Long-term Recurrent Convolutional Networks

CTT-MMC - Convolution-Through-Time for Multi-label Movie genre Classification

TF-IDF - N-grams with Term Frequency–Inverse Document Frequency

LSTM - Long Short-Term Memory

S4 - Structured State Space Sequence

E2ECNN - End-to-End Convolutional Neural Network

TDNN - Time-Delay Neural Network

YOLO-CNN - You Only Live Once Convolutional Neural Network

CBOW - Continuous Bag Of Words

FFMPEG - Fast Forward Moving Picture Experts Group

List of Figures

Chapter 1

Introduction

This chapter comprises of a brief introduction towards the topic at hand, Movie Censorship. The following contents will discuss the field of censorship, exact problem which the project will be tackling, reasons for undertaking this project, expected goals, challenges faced, assumptions, the relevance of this project and an outline of this report's contents.

1.1 Background

The relationship between technology and content control has become essential in the ever-changing world of film. The proliferation of digital platforms and streaming services underscores the critical need for strong filtering procedures to protect audiences from potentially objectionable information. This preface explores how the incorporation of state-of-the-art automation modules is changing the movie censorship scene and provides an insight into the revolutionary potential of artificial intelligence in content regulation.

First, we deploy a highly advanced Human Action Recognition Module as our first step into the world of automated censorship. This module analyses movie sequences by taking individual frames' spatial data and video sequences' temporal features, using a two-stream Convolutional Neural Network (CNN) architecture. The outcome is an accurate and nuanced categorization of human behaviour, which represents a major advancement in the detection of minute details in information that might elude conventional filtering techniques.

In addition, the system has an Inappropriate Content in Image Detection Module that uses a real-time censorship algorithm based on YOLO. This algorithm quickly classifies images and frames in movies as appropriate or inappropriate by using a pipeline-based

design that parallelizes operations with sub processes. This algorithm's real-time nature makes the system more responsive and guarantees a timely and efficient approach to content regulation. When combined, these components create a complex network of technological power that redefines the limits of film censorship and establishes new benchmarks for digital era content monitoring and control.

Our automated censoring system not only analyses visual and spatial content, but also audio dimensions using the Inappropriate Audio Content Module. This package trains an end-to-end CNN model by using a large dataset of improper speech material and Log-Mel spectrograms for feature extraction. The module effectively detects and filters inappropriate audio content by optimising predictions and posterior processing. This adds a critical layer to the whole automation of content oversight. We are going to travel towards a cinematic experience that is not only immersive but also considerate of the varying sensitivities of the audience as we investigate the merging of these advanced modules.

1.2 Problem Definition

There is a pressing demand for an automated censorship application to address the increasing instances of profanity and inappropriate content online. This tool should not only detect and censor offensive material but also classify movie content accurately, providing users with a safer and more personalized digital experience.

1.3 Scope and Motivation

The project's scope includes creating a comprehensive, automated system for movie censoring that makes use of cutting-edge technologies to improve content regulation. Through the integration of modules including human action recognition, real-time image content analysis, audio content monitoring, and genre classification, the project seeks to solve the shortcomings of conventional manual censoring techniques. With real-time replies to guarantee a prompt and precise censorship procedure, this technology aims to offer a sophisticated and effective method of recognising objectionable content in films. Its reach include addressing the dynamic terrain of digital media, tolerating a wide range

of genres and content formats on several platforms, and ultimately making the cinematic experience safer and more pleasurable for viewers.

The urgent need for a technologically sophisticated and automated method of movie censoring is what inspired this initiative. The exponential expansion of digital media and the variety of content channels it offers has made manual content policing more difficult and time-consuming. The goal of this project is to employ cutting-edge technologies, like Log-Mel spectrograms for inappropriate audio content detection, YOLO-based real-time censorship algorithms for inappropriate content in images, and two-stream CNN architecture for human action recognition, to address these challenges. In addition to offering a more effective and precise method of content monitoring, the integration of these modules shows a dedication to changing with the cinematic content landscape and guaranteeing a safer and more pleasurable experience for viewers everywhere.

1.4 Objectives

1. Develop a comprehensive multimodal content analysis system
2. Accurately identify and classify inappropriate content in videos, audio recordings, and images
3. Leverage deep learning and natural language processing techniques
4. Achieve state-of-the-art performance on benchmark datasets

1.5 Challenges

1. Accuracy and False Positives: A major problem is to achieve high content categorization accuracy without producing too many false positives. To avoid needless filtering or missing objectionable content, accuracy and recall must be balanced.
2. Real-time Processing: Ensuring real-time video content processing can be difficult, particularly for the YOLO-based real-time censorship method. Maintaining seamless censoring without sacrificing efficacy requires finding a balance between speed and accuracy.

3. Adaptability to Diverse Content: It is difficult to manage the heterogeneous nature of films, which span several genres, languages, and cultural quirks. The system must be flexible enough to accommodate a wide range of content attributes.

4. Ethical Considerations: It can be difficult to strike a balance between censorship and the right to free speech, particularly when dealing with subjective issues. It is a difficult challenge to provide ethical content regulation without unwarranted censorship or interference with artistic expression.

5. User Experience: It is difficult to implement censorship effectively without negatively affecting the user experience. In order to maintain content regulatory requirements and minimise disturbance to the viewing experience, the system must be easily integrated into already-existing platforms.

6. Continuous Learning: To guarantee that the system can adjust to changing trends in content and new types of unsuitable content, a mechanism for ongoing learning and model changes is needed.

1.6 Assumptions

1. Data Quality: In order to train the machine learning models in each module, the project presupposes access to high-quality datasets, guaranteeing representative and varied samples of human actions, unsuitable content, and different genres.

2. Computational Resources: It is anticipated that the project will have enough processing power to train intricate neural network models, particularly for the 2D CNNs, YOLO-based algorithms, and two-stream CNN architecture used in the genre categorization module.

3. Real-world Applicability: It is assumed by the project that the automated censoring system will work with a wide range of films and different kinds of content, enabling efficient regulation across genres, languages, and cultural contexts.

1.7 Societal / Industrial Relevance

1. User Protection: By attempting to protect users from potentially harmful and improper content, the initiative is relevant to society. In a time when content is readily available on many platforms, protecting users is essential to creating a responsible and safe online community.

2. Cultural Sensitivity: The system exhibits awareness of cultural diversity by including modules like genre classification. This guarantees that cultural sensitivities and nuances are taken into account when regulating content, fostering a more diverse and culturally sensitive entertainment environment.

3. Content Responsiveness: Our system's real-time capabilities take care of content issues as soon as they arise. Because unsuitable content is quickly detected and blocked, users may enjoy a more responsive and safe watching experience while also supporting a healthy digital culture.

4. Effective Content Moderation: When compared to manual content moderation, our automated censoring system gives content providers and platforms a more scalable and effective solution. Processes for delivering information can be streamlined, operating expenses can be decreased, and content management effectiveness can be increased overall.

5. Adherence to Regulations: A number of content laws and grading schemes apply to the entertainment sector. With the help of our initiative, content providers will be able to comply with these requirements more successfully, reducing the possibility of legal problems and encouraging the responsible distribution of material.

6. Improved User Experience: Our product adds to an improved user experience by offering real-time censoring and precise genre classification. User loyalty can be increased by providing more engaging and personalised viewing experiences through content recommendations based on genre classification.

7. Technological Innovation: The project is in line with the entertainment industry's broader tendencies in technological innovation. Using cutting-edge technologies like

Log-Mel spectrograms, YOLO-based algorithms, and two-stream CNN architectures puts the sector at the forefront of technological innovation.

8. Competitive Edge: By providing a safer and easier-to-use experience, content platforms that utilise our automated movie censorship system obtain a competitive advantage. These platforms are probably going to draw a more discriminating user base that is worried about the propriety and quality of the content.

Ultimately, our project is highly relevant to society and business since it tackles current issues with content regulation, encourages a safer online space, and reflects changing consumer and industry demands.

1.8 Organization of the Report

- Brief explanation of the current situation in the field of genre identification in movies as well as mentions of methods found.
- An estimate of the hardware and software requirements.
- Detailed explanation of methods that can be found used in this project. Involves description of the architecture and a breakdown of the work done for the project.
- Providing insights into the overall knowledge acquired during the undertaking of this project and creation of the report.
- References from which the content of the report is based on.
- All the slides of the project presentation.
- Programme Outcomes, Course Outcomes and their mapping.

The chapter provides an outline as to what the report will entail in the coming pages. This report will focus on information gathering, investigation as well as a document of the progress achieved so far in the development of this project.

Chapter 2

Literature Survey

2.1 Recognition of Posture for Abnormal Behavior Detection

This method [1], proposed by Nar et al., utilizes Kinect 3D camera and logistic regression to recognize posture for identifying unusual activity in front of ATMs. It is particularly useful for enhancing video monitoring systems' performance in real-time and improving retrieval systems' performance for historical videotapes or other media sources.

2.2 Algorithm Using Motion Vector for Violence Detection

Xie et al. proposed [2] an algorithm that removes motion vectors from compressed videos, analyzes space-time distribution, and uses radial-based SVM classification for violence detection in surveillance videos. This method is effective in increasing the accuracy of video monitoring systems for real-time detection of violent activities and improving retrieval systems' performance for identifying instances of violence from old videotapes or other media.

2.3 Automatic Special Violence Detection Technique

Senst et al. introduced [3] a method for modelling long-term temporal structure in violence detection that is based on the Lagrangian method and the FightNet convolutional neural network (CNN). This method is particularly valuable for accurately detecting violent actions in videos and modeling long-term temporal structure for improved violence detection.

2.4 DeepCens: A Deep Learning Based System for Real Time Image and Video Censorship

The discussed approach here [4], proposed by Yuksel et al., employs a deep learning based system for detection of inappropriate content using a two stage pipelining architecture. This uses the YOLO-CNN deep learning model, trained on curated annotated datasets for violence, firearms, substance abuse, nudity and sexual content, and intends a realtime censorship solution. Due to the pipeline approach, the detected content timestamps, duration and frame parts are passed as metadata to a censorship chain, which is then buffered concurrently for blurring the content, promising high accuracy, performance and speed in detection. Comparisons with alternative and contemporary approaches were performed, such as with Single Shot Detector (SDD) and Faster R-CNN methods, over multiple explicit content categories, with the proposed approach offering greater accuracy and lesser processing time.

2.5 Cover the Violence: A Novel Deep-Learning-Based Approach Towards Violence-Detection in Movies

This article [5] introduces a violence detection scheme for movies that consists of three main steps. Firstly, the entire movie is segmented into shots, and a representative frame from each shot is selected based on saliency levels. These selected frames are then passed through a light-weight deep learning model, specifically a fine-tuned MobileNet using transfer learning, to classify shots as either violent or non-violent. Finally, the non-violent scenes are merged in sequence to generate a violence-free version of the movie. The proposed model is evaluated on three violence benchmark datasets, demonstrating fast and accurate detection of violent scenes compared to existing methods.

The authors emphasize the complexity of movie data, which comprises different scenes with shots. Prior to violence detection, preprocessing mechanisms based on hand-engineered features are employed to structure the movie data effectively. Key contributions include segmenting the movie into shots and extracting salient frames to aid in violence detection. The proposed framework aims to automatically detect and cover violent scenes in movies, catering to the increasing accessibility of entertainment resources for children and sensitive individuals through smart devices.

Furthermore, the authors fine-tune a light-weight deep CNN model (MobileNet) using pre-trained ImageNet weights to enhance violence recognition in datasets. By leveraging deep learning techniques, the proposed approach addresses the computational expense associated with violence recognition methods in the literature. The model's ability to converge easily and accurately detect violent scenes in movies contributes to a more efficient and automated system for classifying movie content based on violence levels.

2.6 Filtering of Inappropriate Video Content: A Survey

Taha et al. [6] provides a comprehensive overview of the methods and challenges in filtering sensitive content from videos, particularly focusing on issues like violence, pornography, and inappropriate material for children, and discusses the benchmarking and filtering of inappropriate video content through various methods and criteria, based on approach, both machine learning and non-machine learning based, based on type of content, such as bodily harm, violence, substance abuse, and sexual content, and based on media fidelity, animated content against real-life content, and data types for classification. It also considers the performance evaluation for the reported methods.

2.7 Other Works

The following table provides a lot of other works related to one or more modules of the project.

Paper Name	Author	Abstract
Two-Stream Convolutional Networks for Action Recognition in Videos	Karen Simonyan Andrew Zisserman	A two-stream ConvNet architecture which incorporates spatial and temporal networks
Towards Closing the Energy Gap Between HOG and CNN Features for Embedded Vision	Amr Suleiman, Yu-Hsin Chen, Joel Emer, Vivienne Sze	We provide an in-depth analysis of deep Convolutional Neural Networks (CNN) and Histogram of Oriented Gradients (HOG)
Human activity detection and action recognition in videos using convolutional neural networks	Jagadeesh Basavaiah & Chandrashekar Mohan Patil	Optical Flow familiarization
Faster Human Activity Recognition with SVM	K. G. Manosha Chathuramali, Ranga Rodrigo	Advantages of SVM used as final classifier
Fusion of histogram based features for Human Action Recognition	Suraj Prakash Sahoo, Silambarasi R, Samit Ari	Using HOG and BoHOG for motion detection

Figure 2.1: Existing Works

Filtering of Inappropriate Video Content A Survey	Eng. Mahmoud Mohammed Taha, Prof. Abdel Wahab Alsammak, Dr. Ahmed B.Zaky	Deep learning for filtering of inappropriate content using SSD MultiBox and YOLO
Transfer Detection of YOLO to Focus CNN's Attention on Nude Regions for Adult Content Detection	Nouar AlDahoul, Hezerul Abdul Karim, Mohd Haris Lye Abdullah, Mohammad Faizal Ahmad Fauzi, Abdulaziz Saleh Ba Wazir, Sarina Mansor and John See	YOLO CNN models comparison and score understanding for explicit content; on ResNet101 with random forest approach
Cover the Violence: A Novel Deep-Learning-Based Approach Towards Violence-Detection in Movies	Samee Ullah Khan, Ijaz Ul Haq, Seungmin Rho, Sung Wook Baik and Mi Young Lee	Detection of violence scene using Transfer Learning Approach in CNN (MobileNet)
Generative Adversarial Nets	Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio	GAN model for Obscuring inappropriate content
A Benchmarking Campaign for the Multimodal Detection of Violent Scenes in Movies	Claire-H´el‘ene Demarty1, C´edric Penet1, Guillaume Gravier, and Mohammad Soleymani	Multimodal detection of violence scenes using supervised classification systems

Figure 2.2: Existing Works

Deep Learning based Detection Of Inappropriate Speech Content For Film Censorship	ABDULAZIZ SALEH BA WAZIR , HEZERUL ABDUL KARIM , (Senior Member, IEEE), HOR SUI LYN, (Student Member, IEEE), MOHAMMAD FAIZAL AHMAD FAUZI , (Senior Member, IEEE), SARINA MANSOR , AND MOHD HARIS LYE ,	An advanced deep learning system, using CNNs and Log-Mel spectrograms, to improve automated detection of inappropriate speech in films
Small-footprint Keyword Spotting Using Deep Neural Networks	Guoguo Chen, Carolina Parada, Georg Heigold	combining deep neural networks and Hidden Markov Models (HMM), achieving significantly improved recognition accuracy, particularly in noisy conditions
Adversarial Examples For Improving End-to-end Attention-based Small-footprint Keyword Spotting	Xiong Wang, Sining Sun, Changhao Shan, Jingyong Hou1, Lei Xie, Shen Li, Xin Lei	Focuses on improving keyword spotting with adversarial examples, leading to significant performance enhancements in speech recognition.

Figure 2.3: Existing Works

Improving RNN Transducer Modeling For Small-footprint Keyword Spotting	Yao Tian, Haitao Yao, Meng Cai, Yaming Liu, Zejun Ma	Improves small-footprint keyword spotting by enhancing the RNN-T model, leading to better speech recognition performance.
Spectrogram-based Classification Of Spoken Foul Language Using Deep CNN	Abdulaziz Saleh Ba Wazir, Hezerul Abdul Karim, Mohd Haris Lye Abdullah, Sarina Mansor, Nouar Aldahoul, Mohammad Faizal Ahmad Fauzi, John See	CNN-based model for accurate foul language detection in audio using spectrogram images, achieving high classification performance
A Multimodal Approach For Multi-label Movie Genre Classification	Rafael B. Mangolin, Rodolfo M. Pereira, Alceu S. Britto Jr, Carlos N. Silla Jr, Diego Bertolini	A multimodal approach that takes into consideration movie trailers, synopsis, audio, subtitles, posters using MLCC, C3D, SSD, etc.
Rethinking Genre Classification With Fine Grained Semantic Clustering	Edward Fish, Jon Weinbren, Andrew Gilbert	Uses only visual and audio data to classify movies with the help of pretrained expert networks.

Figure 2.4: Existing Works

Long Movie Clip Classification With State-space Video Models	Md Mohaiminul Islam, Gedas Bertasius	Uses S4 layer to decode features extracted by a Standard Transformer encoder in long form videos
Exploiting Deep Learning And Explanation Methods For Movie Tag Prediction	Erica Coppolillo, Massimo Guarascio, Marco Minici, Francesco Sergio Pisani	Uses a deep learning based hierarchical multilabel classifier that contains 3 layers.
Predicting Genre From Movie Posters	Gabriel Barney and Kris Kaya	Uses ResNet34 and a custom architecture to assign genres to posters.

Figure 2.5: Existing Works

2.8 Summary and Gaps Identified

The chapter provides an insight into the existing architectures that can be found relating to the topic at hand. The works done by other fellow researchers provides information on the different methods, models, architecture, infrastructure, etc. that can be used to create models involving movie censorship.

2.8.1 Gaps

- The existing works mostly focus on single modal projects.
- The available works do not incorporate all the aspects of censorship.
- Genre identification is mainly focused on short form videos.

Chapter 3

Requirements

3.1 Hardware and Software Requirements

3.1.1 Hardware Requirements

- Dedicated CUDA GPU: Minimum Video Memory - 8 GB
- Minimum RAM requirement - 32 GB

3.1.2 Software Requirements

- Whisper ASR
- Kaldi toolkit
- VOSK api
- vosk-model-en-us-0.42-gigaspeech
- Open Neural Network Exchange (ONNX) SDK
- .NET Version: 8.0
- Tensorflow Version: 2.3.0 and newer
- Datasets (Gigaspeech, TMDB, TAPAD, ImageNet, etc.)
- Python Version: 3.5+
- PIP Version: 20.3 and newer.
- Google Colab
- Jupyter Notebook Version: 7.0.6 and newer
- Windows x86/x64

3.2 Functional Requirements (Numbered List/ Description in Use Case Model)

FR-1 User Input:

- The system shall accept an uncensored movie as input from the user.

FR-2 Audio Processing - Inappropriate Speech Detection Module:

- The system shall extract audio from the input movie.
- Spatial and temporal features shall be extracted from the audio accurately.
- The inappropriate words and phrases shall be identified properly.
- A transcript shall be created for the implementation of the genre classification.
- The detected words shall be muted or censored effectively without affecting the rest of the audio content.

FR-3 Video Processing - Image Detection Module:

- The system shall simultaneously process the video stream.
- The system shall accurately identify the various types of inappropriate visual content in the video stream frames accurately and rapidly.
- The detected inappropriate video content shall be blurred effectively and for the necessary duration.

FR-4 Video Processing - Human Action Recognition Module:

- The system shall identify the relevance of the detected action.
- Relevancy of action in the movie and dependence of inappropriate content shall be established.

FR-5 Genre Classification Module:

- The system shall accurately classify the genre of the given movie using the transcripts obtained from the audio censorship module.
- The input movie shall have a corresponding relation to the censorship applied.

FR-6 Censorship Module:

- The outputs from the Inappropriate Speech Detection Module, Image Detection Module, Human Action Recognition Module, and Genre Classification Module shall be combined.
- The Censorship Module shall orchestrate the final decision-making process, applying various censorship techniques such as blurring or muting based on detected inappropriate content.
- The censored version of the movie shall be securely stored and its metadata as a censorship chain obtained.

FR-7 Output:

- The final output of the system shall be the censored version of the movie, refined and compliant with predefined censorship standards.

Chapter 4

System Architecture

The system architecture for the proposed approach for movie censorship, MoRedact, is discussed in detail in this chapter. The architectural design and module division in this approach are analyzed.

4.1 System Overview

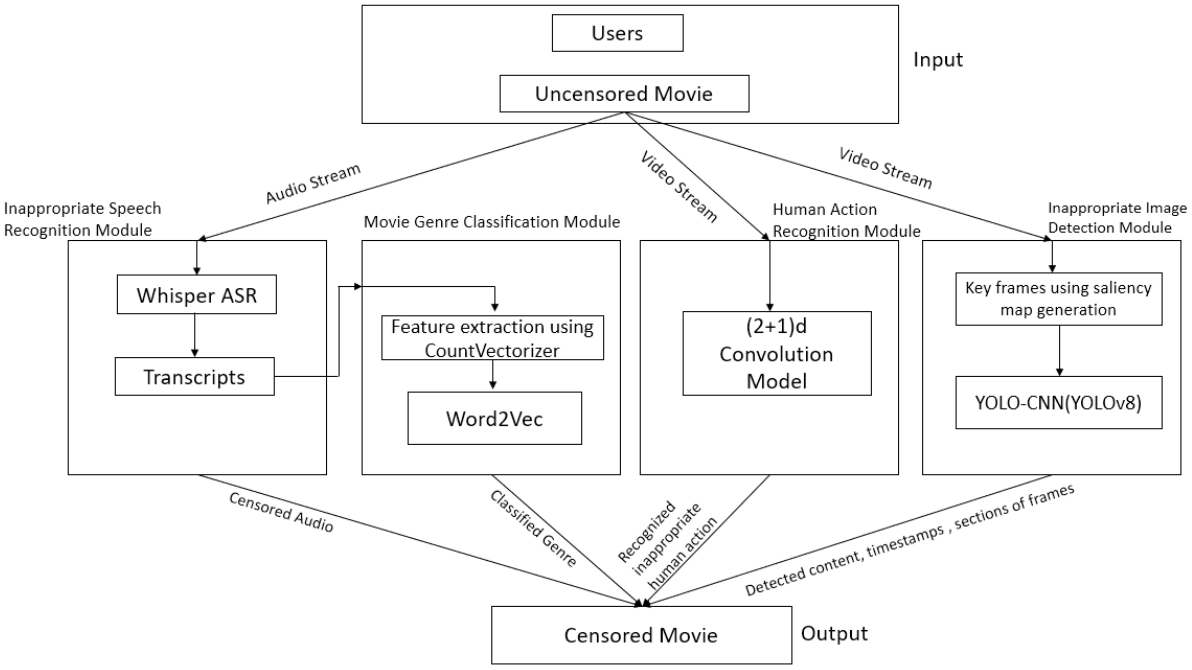

Figure 4.1: Architecture Diagram

4.1.1 Input

This section comprises of the end users and their inputs. The inputs to this detection and censorship system consist of the uncensored movie, where the movie would be passed

to the audio and video processing modules, and the audio transcript is further used for the genre classification module.

4.1.2 Audio Processing - Inappropriate Speech Detection Module

Audio is extracted from the input content as a .WAV file. The audio file undergoes a process of speech to text conversion, identification of timestamps of each word, identification of words to be censored and replacement of the designated words with the sound nullified.

4.1.3 Video Processing - Image Detection Module

The module consists of four stages:

1. Decoding: Algorithms depending on the video format are used.
2. Computation: Processing of video data using DL models (YOLO-CNN model - YOLOv8) with methods to decrease total time taken (saliency and pipelining)
3. Metadata: Creation of censorship chain for each movie frame which contain timestamps and location of harmful content.
4. Encoding: Ensure that the processed video data is re-compressible and storable.

4.1.4 Video Processing - Human Action Recognition Module

The approach takes visual media as its input, which are separated into individual frames that undergo normalisation. The training data consists of safe and unsafe videos, which are validated using labels corresponding to the same respectively. A combination of Convolutional 3D layers and MaxPooling 3D layers are used to extract features relevant for classification and reducing the dimensions of feature maps respectively.

4.1.5 Genre Classification Module

In parallel to the censorship module, a genre identification module works using the transcripts created by the audio censorship module as its input. The module utilises TF-IDF as its method, and works by following a Word2Vec architechture [7].

The dataset is sourced from The Movie DataBase (TMDB), which is an open source movie information and metadata database.

4.1.6 Censorship Module

The outputs from the Inappropriate Speech Detection Module, Image Detection Module, Human Action Recognition Module, and Genre Classification Module are combined. This combined information is processed in the Censorship Module for the final decision-making process. Various censorship techniques, such as blurring or muting, are applied based on the detected inappropriate content in both audio and video components. The censored movie is securely stored in a database.

4.1.7 Output

The final output of the system is the censored version of the movie, which adheres to predefined censorship standards, where the detected inappropriate content is blurred according to the timestamp duration and frame sections, and the detected inappropriate speech parts are muted out.

4.2 Architectural Design

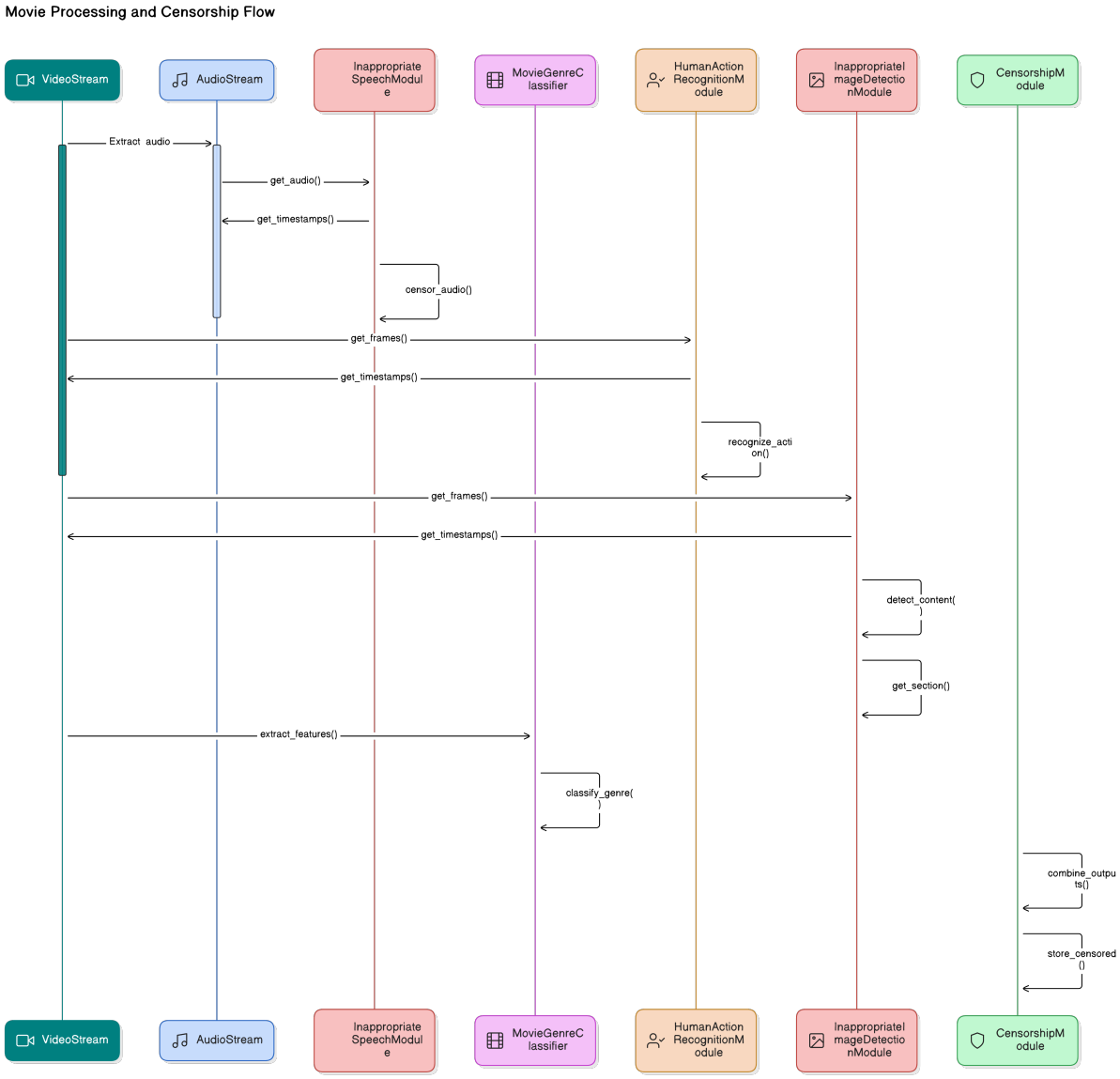

Figure 4.2: Class Diagram

4.3 Module Division

The proposed method is divided into a few separate modules which focuses on each aspect of the censorship process. The uncensored movie as well as the movie posters are sent to the MoRedact system, and the audio and video streams are extracted.

4.3.1 Inappropriate Speech Detection Module

In this project, we are making use of the 'small' Whisper architecture model provided by OpenAI, trained on a custom dataset. The Whisper architecture [8] is a simple end-to-end approach, implemented as an encoder-decoder Transformer.

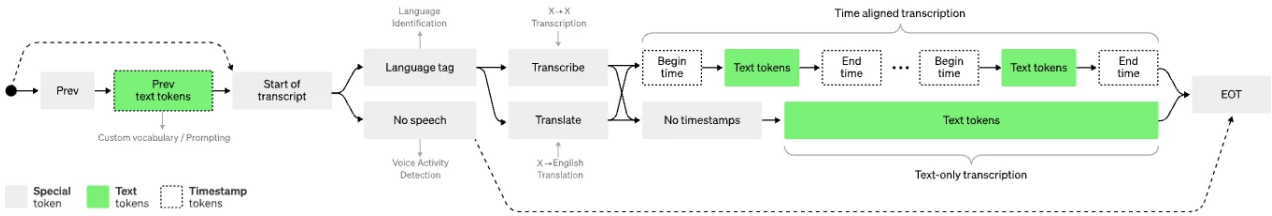

Figure 4.3: Whisper ASR Architecture

The ASR provides both the transcription and timestamps of each word found in the audio. The transcribed text is further used as the input for the genre identification module.

4.3.2 Inappropriate Image Detection Module

The video stream from the uncensored movie is channeled into the Image Detection Module. Here, a pipeline method is implemented, beginning with the division of the video stream into individual frames. Saliency mapping techniques are then applied to identify salient frames, from which key frames are extracted. These key frames are subsequently passed to the YOLO model, trained on a subset of the COCO dataset, for precise object detection for inappropriate content like violence, substance abuse and nudity. Concurrently, blurring of the detected timestamps and locations are performed. This approach was trained on a specially curated subset of the COCO dataset, with emphasis on inappropriate content relating to movies.

4.3.3 Movie Genre Classification Module

The input for the movie genre classification module is in the form of textual data. The audio censorship module, in addition to censoring the audio, also succeeds in creating a highly accurate transcript of the respective audio.

The genre classification follows a Word2Vec architecture [7]. A multi-label genre classification system is developed, utilizing a dataset sourced from The Movie DataBase (TMDB). Following data preprocessing, feature engineering techniques are applied, with TF-IDF selected as a prominent method.

4.3.4 Human Action Recognition Module

The Human Action Recognition (HAR) module performs video classification based on actions depicted in a video. The proposed model is trained on a manually created custom dataset.

The dataset is created based on two labels, content to be censored and content not to be. A total of visual media clips consisting of 788 safe content and 528 unsafe content have been manually collected. The clips are concatenated based on their labels. For training purposes, the concatenated forms are split into parts of 4 seconds each.

A visual media is chosen as the input for the approach, which is then separated into individual frames which are normalised. The training data with bopth labels of safe and unsafe videos are validated with respect to their labels. Feature extraction, relevant for classification and dimension reduction of the feature maps is made possible using a combination of Convolutional 3D layers and MaxPooling 3D layers.

The HAR module follows a (2+1)D convolutional architecture [9]. The process involves decomposing spatial and temporal dimensions. Unlike a 3D convolutional architecture that combines all the vectors from a 3D patch of volume to process time and space dimensions, the (2+1)D convolution processes the time and space dimensions separately, reducing the amount of parameters required to less than half of what was originally required when using 3D convolutional architecture.

4.3.5 Module-wise Division

- Ameen Mohammed: Human Action Recognition
- Ankit John: Inappropriate Image/Content Detection
- Adithya M: Inappropriate Speech Detection

- Ajith Bobby: Movie Genre Classification

4.4 Work Schedule - Gantt Chart

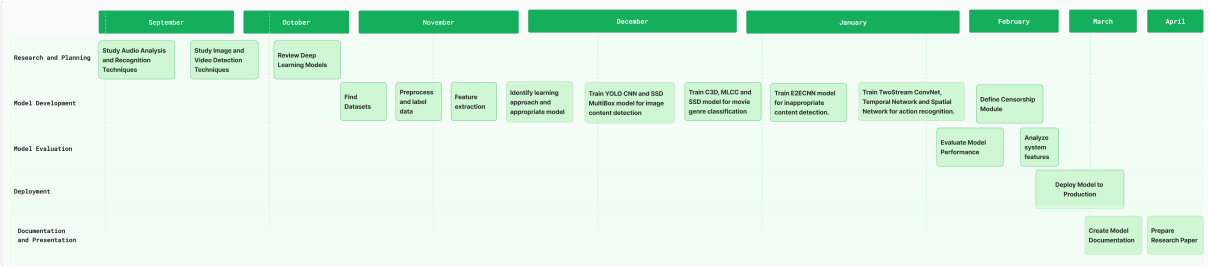

Figure 4.4: Gantt Chart

4.5 Conclusion

This chapter deals with the system overview for the proposed approach to movie censorship, MoRedact, and delves into detail into the system architecture and class diagram for the same. The modules involved in this detection and censorship system as well as the module wise work division has also been discussed, with emphasis on the novelty of the proposed approach.

Chapter 5

System Implementation

This chapter deals with the system implementation of the proposed strategy towards comprehensive video censorship.

5.1 Datasets Identified

5.1.1 Audio Censorship

The dataset used for fine-tuning the speech recognition model for obtaining the transcripts is the GigaSpeech dataset [10], which is a Multi-domain ASR Corpus with 10,000 hours of transcribed audio.

5.1.2 Genre Classification

For training of the genre classification models, a custom dataset of 1000 movies was created with movie information scraped from The Movie DataBase (TMDB), which is further preprocessed and cleaned for improved results. This dataset was further refined with relevant movies, those with a minimum viewership count and through pairwise analysis of movie genres. Further, proper labelling of the movie data was performed with a standardized list of 20 genre labels.

5.1.3 Human Action Recognition

The dataset was meticulously constructed by curating scenes from the movies, focusing on intense action and emotional conflict. This involved selecting sequences portraying gunfights, hand-to-hand combat, swordplay, and other perilous encounters, all labeled under the category of 'violence'. Initially, 264 of these scenes were gathered. In an effort to enhance diversity within the dataset, each video was mirrored, effectively doubling the number of instances to a total of 528.

The dataset was created with scenes from movies, with action sequences involving guns, fighting, stabbing, critical wounds, and more under the label of 'violence'. Around 264 violent scenes were collected and to increase the count, the videos were mirrored therefore achieving a total count of 528 videos for violence class.

For non violence and random action videos, from the UFC-101 dataset [1] were chosen and compiled under the label of non violence. Around 799 videos are present combined in the respective classes.

5.1.4 Inappropriate Image Detection

For Violence, bodily harm and weapons, the datasets[2] used for training were created with 350 video clips labelled as "non-violent" and "violent", where non-violent clips are specifically recorded to include behaviours that can cause false positives in the violence detection task, due to fast movements and the similarity with violent behaviours, for improved performance.

For sexual content and nudity, the NPDI-2k Dataset, which consists of around 2000 videos of violent and non-violent content in modern movies with proper annotations and the Large-Scale Pornographic Dataset (LSPD) [11], which contains 500,000 images and 4,000 videos, with more than 50,000 annotated images.

For substance abuse, training was performed with datasets sourced from the Common Objects in Context Dataset (COCO) [3], where relevant actions such as smoking cigarettes, alcohol usage, drug usage through inhalation, injections, etc., were identified as subsets and combined to obtain the object classes for class probability predictions.

Each of the images and video frames were clearly labelled according to the desired explicit content classes, and annotated as such for their bounding boxes.

5.2 Proposed Methodology/Algorithms

5.2.1 Audio Censorship

The audio censorship module was implemented using two ASR architectures, Kaldi ASR trained on the Gigaspeech dataset, implemented using the VOSK api and the Whis-

[1] https://www.crcv.ucf.edu/data/UCF101.php

[2] https://github.com/airtlab/A-Dataset-for-Automatic-Violence-Detection-in-Videos

[3] https://cocodataset.org

per ASR developed by OpenAI. The Whisper ASR was chosen as the appropriate model due do convenience in adaptability with the python script for audio manipulation and creation of transcripts.

Kaldi ASR

In the Kaldi model[4], the audio from the uncensored video is extracted for detailed analysis. Spatial and temporal features are extracted using Mel-Frequency Cepstral Coefficients (MFCC). The speech recognition part was implemented using a Time-Delay Neural Network (TDNN) with Nnet3 architecture. The model was trained on the Gigaspeech dataset [10].

The model was passed through a Speech Recognition wrapper, which provided accurate transcriptions, but due to compatibility issues with the script utilised for censoring and audio manipulation, the model was abandoned.

Whisper ASR

In this project, we are making use of the 'small' Whisper architecture model provided by OpenAI, trained on a custom dataset. The Whisper architecture [8] is a simple end-to-end approach, implemented as an encoder-decoder Transformer. The audio is given as input after splitting it into 30 second chunks, which are converted into their log-Mel spectrogram and encoded afterwards. To predict the respective text caption, a decoder is trained, mixed with specialised tokens that instructs the single model to perform language identification, multilingual speech transcription, phrase-level timestamps and to-English speech translation [5].

Audio is extracted from the input content as a .WAV file. The audio file undergoes a process of speech to text conversion, identification of timestamps of each word, identification of words to be censored and replacement of the designated words with the sound nullified. The identification of words and transcription of the audio file is done by a python script that utilises the capabilities of Whisper ASR. The processing of the input

[4]https://github.com/kaldi-asr/kaldi/tree/master/egs/gigaspeech/s5

[5]https://huggingface.co/openai/whisper-large

file is done using the Fast Forward Moving Picture Experts Group (ffmpeg) [6] project library. The audio track of the input file is extracted, processed and run through the Whisper ASR model, to obtain the transcription. The transcription is obtained as a word by word format as detected in the speech. Each word is assigned its corresponding time frame (start and end) during which it is heard. The timestamps are used in nullifying the specific points of time during which an explicit word might be said.

The censored audio is merged with the censored video in the final stage of the application. The transcription is also extracted as a text document, which is used as the input for the genre classification.

5.2.2 Genre Identification

For the custom multi-label dataset for genre classification, pairwise analysis of movie genres is performed, by first defining a function makes all possible pairs from of genres. Then, we pull the list of genres for a movie and run this function on the list of genres to get all pairs of genres which occur together. This is shown in 5.4.

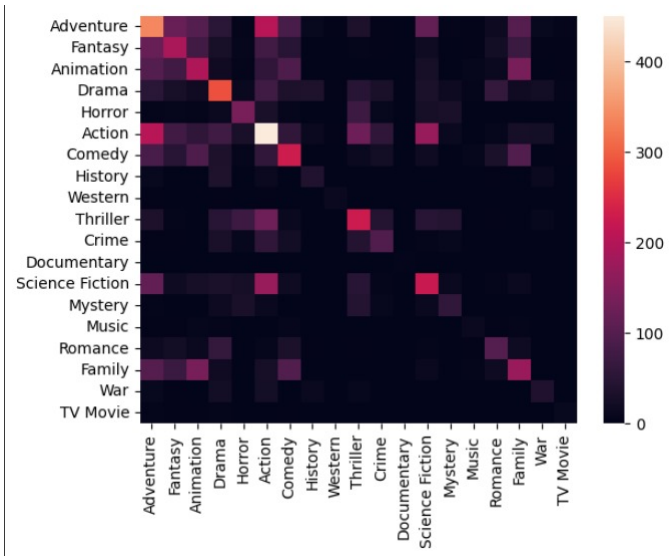

Figure 5.1: Before Biclustering

Then bi-clustering of the obtained pairs is performed to logically determine areas of high intensity (genres that occur frequently with some other specific genres) and overlapping genres. This is shown in 5.2.

[6] https://ffmpeg.org/

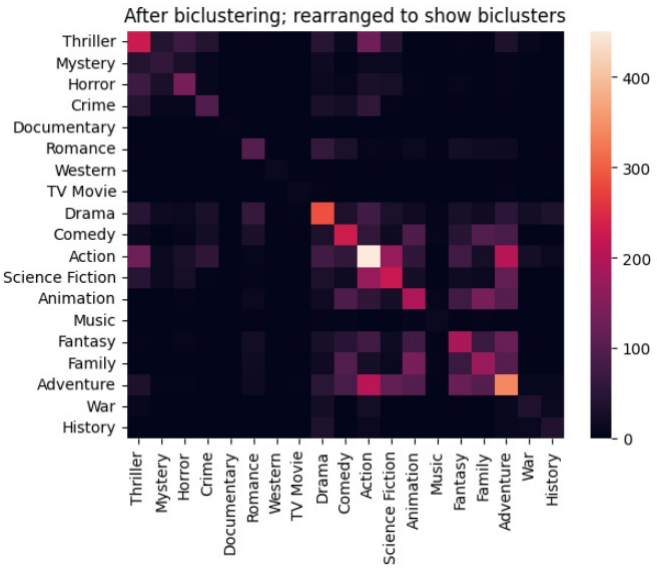

Figure 5.2: After Biclustering

TF-IDF calculates the significance of a word in a document relative to a corpus, emphasizing words that are frequent in a particular document but not across the entire dataset, thus capturing unique textual characteristics for each movie.

Regarding text representation, the raw textual data undergoes conversion into a format suitable for machine learning algorithms. Stop words are eliminated to reduce noise and enhance the efficiency of subsequent analyses. The text is then transformed into numerical format using Word2Vec, a widely used technique for word embedding. Word2Vec represents words as dense vectors in a continuous vector space [12], capturing semantic similarities between words. This is achieved through two main architectures: Continuous Bag of Words (CBOW) and Continuous Skip-gram.

In CBOW, the current word is predicted based on its context, formed by surrounding words in a fixed-size window. This architecture is effective for smaller datasets and frequent words. In contrast, Skip-gram predicts surrounding words given the current word, making it suitable for larger datasets and infrequent words[7]. Both architectures are trained using a neural network to optimize the embeddings, adjusting vector representations to effectively capture semantic relationships.

[7]https://github.com/Spandan-Madan/DeepLearningProject

By converting movie transcripts into 300-dimensional vectors using Word2Vec, dense representations preserving semantic information are created. These vectors serve as inputs to the textual features model, likely a machine learning classifier trained on TF-IDF features. This enables multi-label genre predictions with an accuracy of approximately 85%.

5.2.3 Human Action Recognition

Concurrent to the audio processing and genre classification, the visual content is also being processed via the inappropriate content in image detection and Human Action Recognition models.

The (2+1)D CNN model, trained on the custom dataset, is utilized here for human action recognition. For the input video, the visual content is analyzed by performing inference, for violent or non-violent scenes. Once such a sequence of frames is detected as having violent action, its timestamp tuple (beginning timestamp, end timestamp, duration) is passed to a shared buffer. This allows for concurrent processing with the explicit content detection with the YOLOv8 model, and blurring through the pipeline architecture.

5.2.4 Inappropriate Content in Image Detection

The object detection model used is YOLOv8, trained on the custom labelled dataset of explicit content classes. Here, after the video portion has been passed to the Human Action Recognition module and a buffer of timestamp durations has been obtained, due to the proposed pipelining architecture, processing on the same video content is performed by the YOLOv8 model simultaneously.

Once a timestamp tuple has been buffered, where it consists of the beginning timestamp and end timestamp, the saved YOLOv8 model utilizes that portion of video as the input by first segmenting the video into its constituent frames, then a representative frame from each shot is selected based on saliency levels (maximum number of non-zero pixel count). Then the YOLOv8 model runs inference to detect the explicit content classes (violence, bodily harm, firearms, substance abuse, nudity, sexual content), identifies their

bounding boxes and object perimeter, and the detected objects are then blurred through the use of an applied blurring filter.

In the situation where the buffered timestamp tuple doesn't include any explicit content to be blurred, or rather, no explicit content objects are detected through YOLOv8 inference, the entire sequence is then skipped.

In any case, after detection and censoring by the YOLOv8 model, the timestamp tuple is then removed from the buffer, and the process is repeated as long as the buffer is not empty (as long as there is still visual content left to be processed)

5.3 User Interface Design

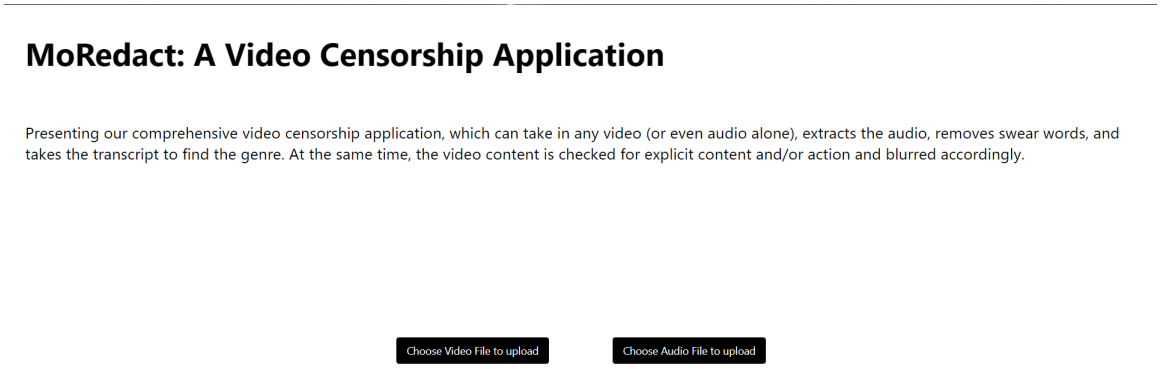

Figure 5.3: Initial User Interface Design

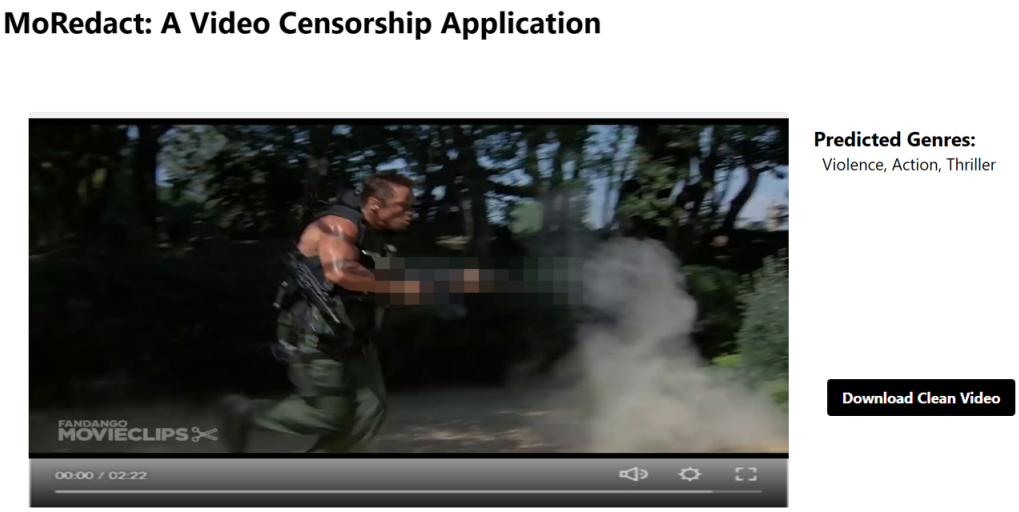

Figure 5.4: User Interface while processing of video and genre classification

5.4 Description of Implementation Strategies

As for the user interface and project implementation strategy, we have decided to go with a React frontend and Flask backend application, where the end user uploads a video for censoring to the application, and would obtain the resultant clean video, without swear words and blurred explicit content, and also the classified genres as well.

After the video upload, an FFMPEG library function is utilized to store the video temporarily, obtain relevant metadata such as bitrate and length of file, and then for extraction of the audio separately as a .AAC file. The custom trained Whisper ASR model is then run on this audio file, to obtain the audio transcript, which is then checked with the pre-defined swear word list, to identify the words to be censored. The corresponding tuple timestamps of the identified swear words and taken, and the volume of the tuple time value is manipulated to be zero via FFMPEG library function. Thus the clean transcript and the clean audio are both acquired.

The derived clean transcript is then passed to the genre classification module, which performs TF-IDF for feature extraction and Word2Vec for embedding on the transcript to convert to a 300 dimensional vector format, and the saved machine learning model is then used to run inference, to attain the three most probable classified genres.

Simultaneously, the video portion is passed to the Human Action Recognition module, which identifies explicit and inappropriate content and the associated timestamps, stored onto a buffer in a pipelining architecture. This buffer is passed to the saved YOLOv8 object detection model, where only the timestamp durations are checked for content to be blurred; if identified, a blurring filter is applied on only the object inside its bounding box. Then the clean, censored visual content is achieved.

Finally, through the use of an FFMPEG library function, the clean audio and visual portions are merged together and displayed to be played back or downloaded, along with the predicted genres, The functionality of user profiling and past history, along with user statistics is also possible.

5.4.1 Summary

The chapter has explained in detail regarding

- the procurement and creation of datasets used
- the methodology followed in the creation of the project, explaining each of the modules and its working in detail
- wireframe designs of the application
- detailed description of the strategies used in the implementation of the system

Chapter 6

Results and Discussions

This chapter focuses on the data obtained, the actions undertaken and the results obtained throughout the creation of this project.

6.1 Overview

The MoRedact project in its final implementation provides an interface through which the end user can input a movie/visual media that they want censored. The video undergoes a separation of audio from it, which goes through the audio censorship module. the video acts as the input for the HAR module as well as the image censorship module. From the audio censorship module, a transcript of the audio is taken to generated a prediction on the genre of the media given by the user.

6.2 Testing

The various modules provide for different results based on its domain. The image censorship provides a blurring of sections of the frame where explicit content specified by the parameters is found. The blurring method used is the classic mosaic style.

The audio censorship provides a visualisation of the words identified in the video given as input, and also depicts the words to be censored in a separate list. The censored audio is then merged with the censored piece of video.

The genre classification utilises the textual data created by the audio censorship module to create a multilabel prediction of the genre to which the input media may belong.

6.2.1 Audio Censorship

The figure shown below shows the textual data obtained from the audio extracted from a sample video given as an input. the words are identified along with the time frames in which they occur, facilitating for audio manipulation in the required timestamps.

```
Word:  avanzar, Start: 3.52, End: 3.96
Word:  Thunder, Start: 3.86, End: 3.96
Word:  Buddy's, Start: 3.86, End: 4.22
Word:  For, Start: 4.12, End: 4.319999999999999
Word:  Life,, Start: 4.22, End: 4.4799999999999995
Word:  right, Start: 4.44, End: 4.68
Word:  Jenny?, Start: 4.58, End: 4.88
Word:  Fucking, Start: 4.92, End: 5.239999999999999
Word:  Fucking, Start: 4.92, End: 5.239999999999999
Word:  Right, Start: 5.14, End: 5.5
Word:  Alright,, Start: 5.4, End: 5.819999999999999
Word:  come, Start: 5.8, End: 5.9799999999999995
Word:  on,, Start: 5.88, End: 6.119999999999999
Word:  let's, Start: 6.04, End: 6.26
Word:  sing, Start: 6.16, End: 6.38
Word:  the, Start: 6.28, End: 6.52
Word:  Thunder, Start: 6.42, End: 6.699999999999999
Word:  Song, Start: 6.6, End: 6.96
Word:  Alright, Start: 6.86, End: 7.319999999999999
Word:  When, Start: 7.22, End: 7.88
Word:  you, Start: 7.78, End: 8.139999999999999
Word:  hear, Start: 8.04, End: 8.379999999999999
Word:  the, Start: 8.28, End: 8.56
Word:  sound, Start: 8.46, End: 8.84
Word:  of, Start: 8.74, End: 9.02
Word:  thunder,, Start: 8.92, End: 9.34
Word:  don't, Start: 9.42, End: 9.76
Word:  you, Start: 9.66, End: 9.879999999999999
Word:  get, Start: 9.78, End: 10.1
Word:  too, Start: 10.0, End: 10.28
Word:  scared, Start: 10.18, End: 10.74
Word:  Just, Start: 10.64, End: 11.28
Word:  grab, Start: 11.18, End: 11.54
Word:  your, Start: 11.44, End: 11.719999999999999
Word:  Thunder, Start: 11.62, End: 12.06
Word:  Buddy, Start: 11.96, End: 12.56
Word:  And, Start: 12.46, End: 13.12
Word:  say, Start: 13.02, End: 13.34
Word:  these, Start: 13.24, End: 13.6
Word:  magic, Start: 13.5, End: 13.959999999999999
```

Figure 6.1: List of all Words identified with timestamps

The figure below shows the words to be censored, which is defined by the user.

```
Iterating through swear word list and muting...
##########
Swear tuple: (' Fucking', 4.92, 5.239999999999999)
Swear tuple: (' Fuck', 14.36, 15.24)
Swear tuple: (' suck', 17.04, 17.360000000000003)
Swear tuple: (' dick', 17.5, 17.880000000000003)
##########
Muting all F-words...
```

Figure 6.2: Identified Word Tuples to be censored

6.2.2 Image Censorship

The following images shows a before and after comparison of the censorship done by MoRedact. The topics of censorship can range from violence, substance abuse, nudity, etc.

Figure 6.3: Example of substance abuse: A man smoking a cigarette

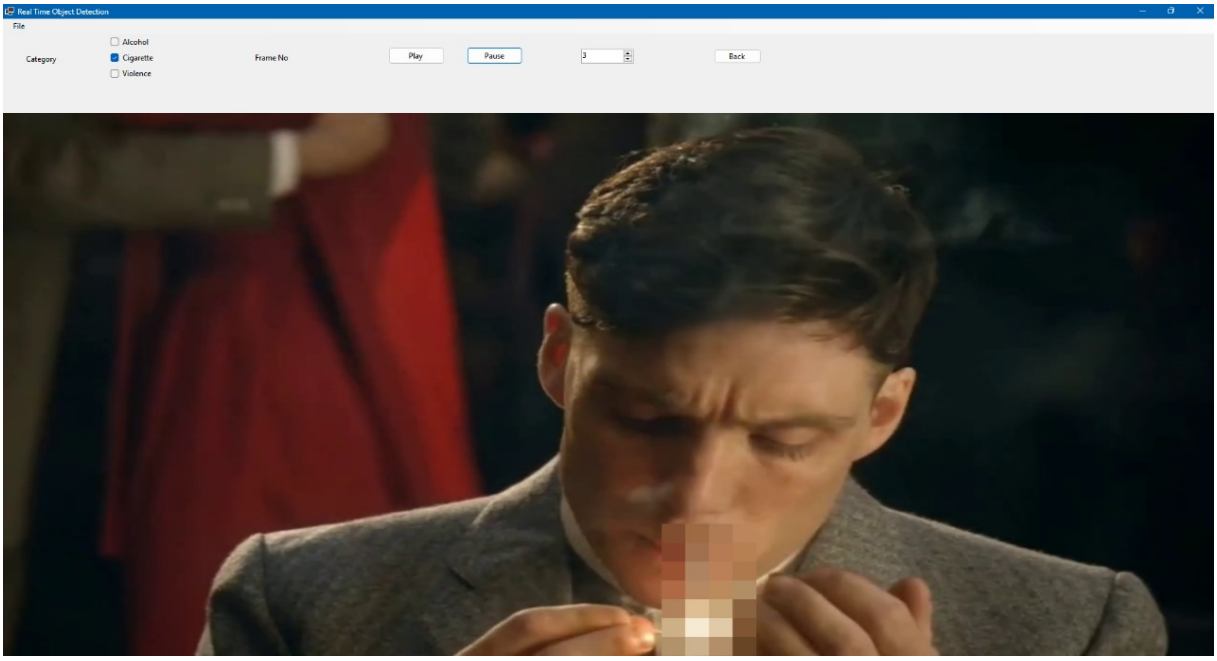

Figure 6.4: After the censorship/blurring of the cigarette

Figure 6.5: Example of an image depicting gun violence

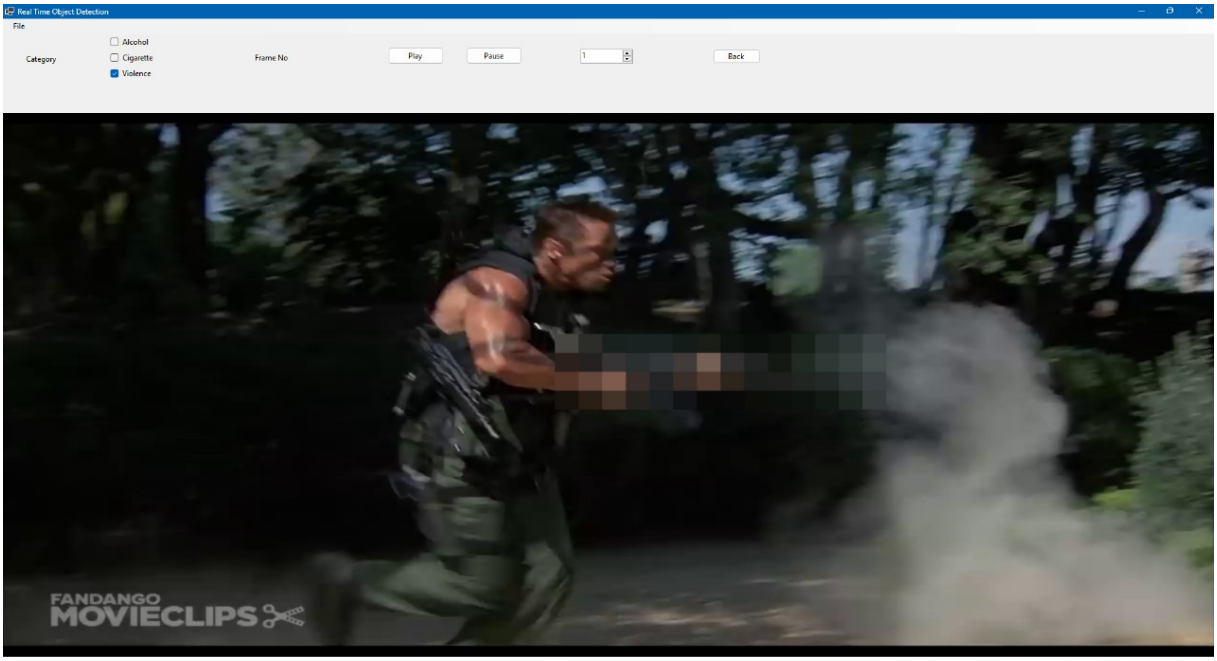

Figure 6.6: After the censorship/blurring of the gun

6.2.3 Genre Classification

The prediction of genre into three labels based on text obtained from transcripts is given in 6.7

```
model_textual = load_model('model_textual.h5')
input_text = "Ana, a college student, interviews an enigmatic billionaire entrepreneur, Christian, for her campus' periodical. A steamy sadomasochistic affair starts between the two, whose roots lie in his past
preprocessed_text = preprocess_text(input_text)
input_vectors = text_to_vectors(preprocessed_text)
input_vectors = input_vectors.reshape(1, -1)
predicted_genres = model_textual.predict(input_vectors)
top_predicted_genres_indices = np.argsort(predicted_genres[0])[-3:]
predicted_genres_list = [Genre_ID_to_name[genre_list[index]] for index in top_predicted_genres_indices]
print("Predicted genres:", predicted_genres_list)

Tokens: ['ana', 'a', 'college', 'student', 'interviews', 'an', 'enigmatic', 'billionaire', 'entrepreneur', 'christian', 'for', 'her', 'campus', 'periodical', 'a', 'steamy', 'sadomasochistic', 'affair', 'starts',
Stopped Tokens: ['ana', 'college', 'student', 'interviews', 'enigmatic', 'billionaire', 'entrepreneur', 'christian', 'campus', 'periodical', 'steamy', 'sadomasochistic', 'affair', 'starts', 'two', 'whose', 'root
Number of valid tokens: 19
1/1 [==============================] - 0s 53ms/step
Predicted genres: ['Comedy', 'Romance', 'Drama']
```

Figure 6.7: Prediction of Genre Using Textual Data

6.3 Quantitative Results

6.3.1 Audio Censorship

The model selected for audio transcription depended on the accuracy and time trade-off while comparing the global Word Error Rate (WER) and average Latency for the models provided by Whisper architecture.

Table 6.1: Comparisons of Whisper Models

Model	WER	Avg. Latency
whisper-tiny	7.54	.43
whisper-base	5.08	.49
whisper-small	3.43	.84
whisper-medium	2.9	1.5
whisper-large	3	1.96
whisper-large-v2	3	1.98

6.3.2 Human Action Recognition

In the human action recognition module, the loss value is calculated through binary cross-entropy data, found to be 0.30 and, the accuracy is calculated as the ratio of the number of correctly classified samples to the total number of samples in the validation set, found to be 97.3%, for the custom trained model.

6.3.3 Genre Classification

Initially a multi-modal approach was considered, which made use of features extracted from the posters as well as the transcript of the audio obtained for the media. But due to unsatisfactory results involving low accuracy (around 65%). This prompted for an approach which only uses the transcript of the speech found in the video. The features obtained just from transcript outperforms the visual features. The precision and recall values were calculated by finding the Numpy average of the individual values for each sample in test data; the precision is found to be 80.4% and the recall is found to be 84.85%. The accuracy of the trained model is 85.24%.

6.3.4 Inappropriate Image Detection

For inappropriate image detection, different approaches were compared among SSD, Faster R-CNN and YOLO. Initially, only for Cigarette content all three approaches were tested, with batch size of 16 and epoch number of 100. YOLO was found to perform better at prediction, with a higher mean Average Precision (mAP) score.

	mAP		
Category	**SSD**	**Faster R-CNN**	**YOLO**
Cigarette	0.45	0.76	.98
Elements of Violence	-	-	.97
Alcohol	-	-	.97
Explicit Content	-	-	.98

Table 6.2: Comparison of mAP values for different models against inappropriate content categories

On testing predictions against the other inappropriate content categories (Substance Abuse, Violence, Explicit Content), the Precision, Recall and F1 scores were obtained.

	YOLO		
Category	**Precision**	**Recall**	**F1**
Cigarette	.984	.99	.98
Elements of Violence	.96	.938	.94
Alcohol	.96	.973	.96
Explicit Content	.973	.98	.97

Table 6.3: Precision, Recall and F1 scores of YOLO predictions on inappropriate content categories

Only the detected explicit content and its bounding box is blurred. The combined accuracy of YOLOv8 on all explicit classes was found to be 92.15%.

6.4 Graphical Analysis

6.4.1 Human Action Recognition

The confusion matrix obtained during the testing is given in 6.8.

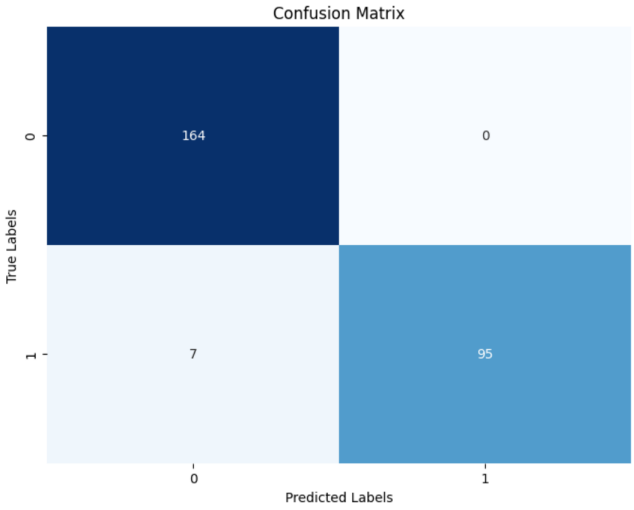

Figure 6.8: Confusion matrix for HAR

The confusion matrix provides a tabular representation of the model's performance, showing the number of correct and incorrect predictions for each class.

1. True Positive (TP): The model correctly predicted 164 instances of class 0 (non-violent videos).

2. False Negative (FN): The model incorrectly predicted 0 instances of class 0 as class 1 (violent videos).

3. False Positive (FP): The model incorrectly predicted 7 instances of class 1 as class 0.

4. True Negative (TN): The model correctly predicted 95 instances of class 1.

Precision, recall and Accuracy

- Precision for class 0 (non-violent videos) is calculated as

$$TP/(TP + FP) = 164/(164 + 7) = 0.96$$

This indicates that out of all the instances predicted as non-violent, 96% were actually non-violent.

- Recall for class 0 is calculated as

$$TP/(TP + FN) = 164/(164 + 0) = 1.00$$

- Accuracy is calculated as

$$(TP+TN)/(TP+TN+FP+FN) = (164+95)/(164+0+7+95) = 259/266 \approx 0.97.$$

This means that the model correctly identified all instances of non-violent videos.

The model performs exceptionally well, with no false negatives for class 0 and a small number of false positives for class 1.

6.5 Discussion

The combination of the four modules have provided a variety of results and insights that can be further studied and improved upon. Individual results obtained have been summarised below :

- The audio censorship module provides a transcript with the timestamp of each word identified. Additionally, it also nullifies the audio at the time frames in which explicit/words to be censored are found.
- The HAR and image censorship together provides an accurate identification of actions in the video, and with the capabilities of YOLOv8, specific parts of the frame where censoring is required can be identified and they are promptly censored.
- The genre classification module predicts the genre of the video content based on the textual analysis of the transcript obtained. The module predicts 3 labels for the input video.

The MoRedact application provides a unified result for the video given as input. The audio part is separated and censored, while creating a transcript of the speech for further prediction of the genre. The video part undergoes action recognition and further identification of explicit content that maybe found throughout the video, which are censored.

The application is implemented as a web application for an efficient and user friendly experience.

Chapter 7

Conclusions & Future Scope

The first phase of the project has provided successful models that work in three modules, independent of each other. The Speech Detection, Image Detection and Movie Genre Classification has been employed. The report outlines the methods used in achieving the said results. The report also encompasses brief descriptions of the existing methods that can be found in this field. The speech detection is employed using speech to text mechanisms, image detection by making use of still images extracted from a video, and mapping features from it. The genre identification module makes use of audio transcripts, benefiting from the textual content found in the dialogues.

The project proposes a novel approach towards a multimodal architecture that aims at providing an efficient model to aid and automate the censorship and classification of movies.

The project aims at furthering its development by incorporating a Human Action Recognition module as well. The model will provide accurate assessment of the action performed in a given instance of the video/movie. After which all the four modules are to be employed in a single architecture. The second phase of the project will focus on the multimodality and its enforcement in the project. The final proposed model being able to censor specified acoustic and visual content as well as classify the movie into its appropriate genre.

References

[1] R. Nar, A. Singal, and P. Kumar, "Abnormal activity detection for bank atm surveillance," in *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2016, pp. 2042–2046.

[2] J. Xie, W. Yan, C. Mu, T. Liu, P. Li, and S. Yan, "Recognizing violent activity without decoding video streams," *Optik - International Journal for Light and Electron Optics*, vol. 127, 11 2015.

[3] T. Senst, V. Eiselein, A. Kuhn, and T. Sikora, "Crowd violence detection using global motion-compensated lagrangian features and scale-sensitive video-level representation," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 12, pp. 2945–2956, 2017.

[4] A. Yüksel and G. Tan, "Deepcens: A deep learning-based system for real-time image and video censorship," *Expert Systems*, vol. 9, p. 0, 08 2023.

[5] S. U. Khan, I. U. Haq, S. Rho, S. W. Baik, and M. Y. Lee, "Cover the violence: A novel deep-learning-based approach towards violence-detection in movies," *Applied Sciences*, vol. 9, no. 22, 2019. [Online]. Available: https://www.mdpi.com/2076-3417/9/22/4963

[6] M. M. Taha, A. B. Zaky, and A. W. Alsammak, "Filtering of inappropriate video content a survey," *International Journal of Engineering Research & Technology (IJERT)*, vol. 11, no. 2, 2022.

[7] N. Fei and Y. Zhang, "Movie genre classification using tf-idf and svm," in *Proceedings of the 2019 7th International Conference on Information Technology: IoT and Smart City*, ser. ICIT '19. New York, NY, USA: Association for Computing Machinery, 2020, p. 131–136. [Online]. Available: https://doi.org/10.1145/3377170.3377234

[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23. JMLR.org, 2023.

[9] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2018, pp. 6450–6459. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00675

[10] G. Chen, S. Chai, G.-B. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang, M. Jin, S. Khudanpur, S. Watanabe, S. Zhao, W. Zou, X. Li, X. Yao, Y. Wang, Z. You, and Z. Yan, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," 08 2021, pp. 3670–3674.

[11] D.-D. Phan, T.-T. Nguyen, Q.-H. Nguyen, H.-L. Tran, K. N. Khoi, Nguyen, and D.-L. Vu, "Lspd: A large-scale pornographic dataset for detection and classification," *International Journal of Intelligent Engineering and Systems*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:245551151

[12] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013. [Online]. Available: http://arxiv.org/abs/1301.3781

Appendix A: Presentation

MoRedact: A Motion-Picture Censorship Application

Ameen Mohammed Adithya M Ankit John Abraham Ajith Bobby

Guide : Mr Harikrishnan M

April 29, 2024

Contents

Problem Definition

Problem Definition

- Develop an automated system for genre classification in movies using deep learning techniques to accurately categorize films into different genres such as action, drama, horror, and others.
- Implement a solution for inappropriate audio and image content censoring in movies by leveraging deep learning algorithms to detect and filter out objectionable audio and visual elements.
- Create a system for human action recognition in movies using deep learning methods to identify and classify various human actions, including violent and non-violent behaviors, to ensure appropriate content filtering.
- Explore the integration of multiple deep learning models to develop a comprehensive solution for movie censorship that encompasses genre classification, inappropriate content censoring, and human action recognition.

Project Objectives

Project Objectives

- Develop a comprehensive multimodal content analysis system
- Accurately identify and classify inappropriate content in videos, audio recordings, and images
- Leverage deep learning and natural language processing techniques
- Achieve state-of-the-art performance on benchmark datasets

Novelty of Idea

Novelty of Idea

Novelty of Idea

- **Advanced Technologies**: Utilizing human action recognition and real-time image analysis.
- **Integration of Modules**: Combining modules for swift and accurate responses.
- **Comprehensive Approach:** Covers various genres and formats on digital platforms.
- **Improving Viewer Experience:** Enhancing safety and enjoyment by regulating objectionable content effectively.

Scope of Implementation

- **Comprehensive System**: Integrating modules for improved content regulation.
- **Real-Time Replies**: Ensuring prompt and precise censorship procedures.
- **Adaptability Across Platforms:** Tolerating diverse content formats on various platforms.
- **Enhancing Safety and Enjoyment:** Making cinematic experience safer and more enjoyable.

Literature Review

Literature Review

Paper Name	Author	Abstract
Two-Stream Convolutional Networks in Action Recognition in Videos	Karen Simonyan Andrew Zisserman	A two-stream ConvNet architecture which incorporates spatial and temporal networks
Towards Closing the Energy Gap between HOG and CNN Features for Embedded Vision	Amr Suleiman, Yu-Hsin Chen, Joel Emer, Vivienne Sze	We provide an in-depth analysis of deep Convolutional Neural Networks(CNN) and Histogram of Oriented Gradients(HOG)
Human activity detection and action recognition in videos using convolutional neural networks	Jagadeesh Basavaiah & Chandrashekar Mohan Patil	Optical Flow Familiarization

Literature Review

Paper Name	Author	Abstract
Faster Human Activity Recognition with SVM	K G Manosha Chathuramali, Ranga Rodrigo	Advantages of SVM used as final classifier
Fusion of histogram based features for Human Action Recognition	Suraj Prakash Sahoo, Silambarasi R, Samit Ari	Using HOG and BoHOG for motion detection
Filtering of Inappropriate Video Content: A Survey	Eng. Mahmoud Mohammed Taha, Prof. Abdel Wahab Alsammak, Dr. Ahmed B.Zaky	Deep Learning for filtering of inappropriate content using SSD MultiBox and YOLO

Literature Review

Paper Name	Author	Abstract
Transfer Detection of YOLO to Focus CNN's Attention on Nude Regions for Adult Content Detection	Nouar AlDahoul, Hezerul Abdul Karim, Mohd Haris Lye Abdullah, Mohammed Faizal Ahmad Fauzi, Abdulaziz Saleh Ba Wazir, Sarina Mansor and John See	YOLO CNN models comparison and score understanding for explicit content; on ResNet101 with random forest approach
Cover the Violence: A Novel Deep-Learning-Based Approach Towards Violence-Detection in Movies	Samee Ullah Khan, Ijaz Ul Haq, Seungmin Rho, Sung Wook Baik and Mi Young Lee	Detection of violence scene using Transfer learning approach in CNN(MobileNet)
Generative Adversal Nets	Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair ,Aaron Courville, Yoshua Bengio	GAN model for Obscuring inappropriate content

Literature Review

Literature Review

Paper Name	Author	Abstract
A Benchmarking Campaign for the Multimodal Detection of Violent Scenes in Movies	Claire-H´el`ene Demarty, C´edric Penet, Guillaume Gravier and Mohammed Soleymani	Multimodal detection of violence scenes using supervised classification systems.
Deep Learning based Detection of Inappropriate Speech Content for Film Censorship	Abdulaziz Saleh Ba Wazir, Hezerul Abdul Karim, Hor Sui Lyn, Mohammed Faizal Ahmed Fauzi, Sarina Mansor and Mohd Haris Lye	An advanced deep learning system, using CNNs and Log-Mel spectrograms to improve automated detection of inappropriate speech in films
Small-footprint Keyword Spotting using Deep Neural Networks	George Chen, Carolina Parada, Georg Heigold	Combining Deep neural networks and Hidden Markov Models(HMM),achieving significantly improved recognition accuracy, particularly in noisy condition

Literature Review

Paper Name	Author	Abstract
Adversarial Examples for Improving End-to-End Attention-based Small-footprint Keyword Spotting	Xiong Wang, Sining Sun, Changhao Shan, Jingyong Hou, Lei Xie, Shen Li, Xin Lei	Focuses on improving keyword spotting with adversarial examples, leading to significant performance enhancements in speech recognition
Spectogram-based Classification of Spoken Foul Language Using Deep CNN	Abdulaziz Saleh Ba Wazir, Hezerul Abdul Karim, Mohd Haris Lye Abdullah, Sarina Mansor, Nouar Aldahoul, Mohammed Faizal Ahmed Fauzi and John See	CNN-based model for accurate foul language detection in audio using spectrogram images, achieving high classification performance
Improving RNN Transducer Modeling For Small-footprint Keyword Spotting	Yao Tian, Haitao Yao, Meng Cai, Yaming Liu, Zejun Ma	Improves small-footprint keyword spotting by enhancing the RNN-T model, leading to better speech recognition performance

Literature Review

Literature Review

Paper Name	Author	Abstract
A Multimodal Approach For Multi-label Movie Genre Classification	Rafael B. Mangolin, Rodolfo M. Pereira, Alceu S. Britto Jr, Carlos N. Silla Jr, Diego Bertolini	A multimodal approach that takes into consideration movie trailers, synopsis, audio, subtitles, posters using MLCC, C3D, SSD etc.
Rethinking Genre Classification With Fine Grained Semantic Clustering	Edward Fish, Jon Weinbren, Andrew Gilbert	Uses only visual and audio data to classify movies with the help of pretrained expert networks.
Long Movie Clip Classification With State-space Video Models	Md Mohaiminul Islam, Gedas Bertasius	Uses S4 layer to decode features extracted by a Standard Transformer encoder in long form videos

Literature Review

Paper Name	Author	Abstract
Exploiting Deep Learning and Explanation Methods for Movie Tag Prediction	Erica Coppolillo, Massimo Guarascio, Marco Minici, Francesco Sergio Pisani	Uses a deep learning based hierarchical multilabel classifier that contains 3 layers
Predicting Genre From Movie Posters	Gabriel Barney and Kris Kaya	Uses ResNet34 and a custom architecture to assign genres to posters.

Methodology

Methodology

- **Audio Censorship**
- Input: .WAV format audio.
- Speech-to-text conversion.
- Word timestamps and censorship identification.
- Replacement of designated words with nullified sound.
- Processing library: Fast Forward Moving Picture Experts Group(ffmpeg)
- ASR model: Whisper (small architecture) by OpenAI.

Methodology

- **Genre Classification**
- Word2Vec architecture used for genre classification.
- TMDB dataset preprocessing.
- TF-IDF feature engineering.
- Word2Vec conversion of textual data to numerical format.
- Creation of 300-dimensional vectors for movie transcripts.
- Multi-label genre prediction with approximately 85% accuracy.

Methodology

- **Human Action Recognition**
- Human Action Recognition (HAR) module performs video classification.
- Training on a custom dataset with labels for safe and unsafe content.
- Concatenation and splitting of clips for training.
- (2+1)D convolutional architecture separates spatial and temporal dimensions.
- Input: individual frames normalized.
- Classification based on extracted features using Convolutional 3D and MaxPooling 3D layers.

Methodology

- **Inappropriate Content Detection**
- Decoding: Algorithms are used according to video format.
- Computation: Processing of video data using DL models (YOLO-CNN model - YOLOv8) with methods to decrease total time taken (saliency and pipelining).
- Metadata: Creation of censorship chain for each movie frame which contain timestamps and location of harmful content.
- Encoding: The processed video data must be made re-compressible and storable.

Architecture Diagram

Architecture Diagram

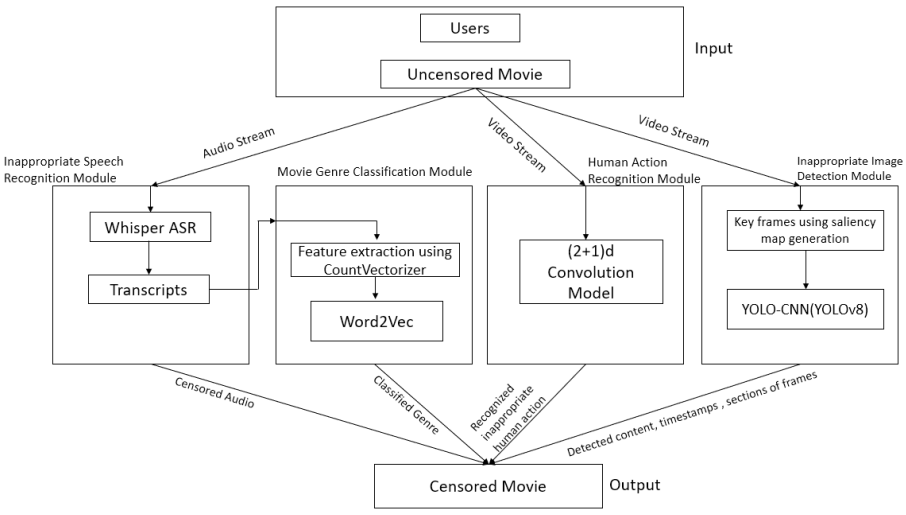

Figure: Architecture Diagram

Work Division

Work Division

Work Division

- **Ameen Mohammed**: Human Action Recognition
- **Adithya M**: Audio Censorship Module
- **Ankit John Abraham**: Inappropriate image detection, Application Interface
- **Ajith Bobby**: Datasets,Genre Identification

Results

Results

```
Word:  avanzar, Start: 3.52, End: 3.96
Word:  Thunder, Start: 3.86, End: 3.96
Word:  Buddy's, Start: 3.86, End: 4.22
Word:  For, Start: 4.12, End: 4.319999999999999
Word:  Life,, Start: 4.22, End: 4.4799999999999995
Word:  right, Start: 4.44, End: 4.68
Word:  Jenny?, Start: 4.58, End: 4.88
Word:  Fucking, Start: 4.92, End: 5.239999999999999
Word:  Fucking, Start: 4.92, End: 5.239999999999999
Word:  Right, Start: 5.14, End: 5.5
Word:  Alright,, Start: 5.4, End: 5.819999999999999
Word:  come, Start: 5.8, End: 5.9799999999999995
Word:  on,, Start: 5.88, End: 6.119999999999999
Word:  let's, Start: 6.04, End: 6.26
Word:  sing, Start: 6.16, End: 6.38
Word:  the, Start: 6.28, End: 6.52
Word:  Thunder, Start: 6.42, End: 6.699999999999999
Word:  Song, Start: 6.6, End: 6.96
Word:  Alright, Start: 6.86, End: 7.319999999999999
Word:  When, Start: 7.22, End: 7.88
Word:  you, Start: 7.78, End: 8.139999999999999
Word:  hear, Start: 8.04, End: 8.379999999999999
Word:  the, Start: 8.28, End: 8.56
Word:  sound, Start: 8.46, End: 8.84
Word:  of, Start: 8.74, End: 9.02
Word:  thunder,, Start: 8.92, End: 9.34
Word:  don't, Start: 9.42, End: 9.76
Word:  you, Start: 9.66, End: 9.879999999999999
Word:  get, Start: 9.78, End: 10.1
Word:  too, Start: 10.0, End: 10.28
Word:  scared, Start: 10.18, End: 10.74
Word:  Just, Start: 10.64, End: 11.28
Word:  grab, Start: 11.18, End: 11.54
Word:  your, Start: 11.44, End: 11.719999999999999
Word:  Thunder, Start: 11.62, End: 12.06
Word:  Buddy, Start: 11.96, End: 12.56
Word:  And, Start: 12.46, End: 13.12
Word:  say, Start: 13.02, End: 13.34
Word:  these, Start: 13.24, End: 13.6
Word:  magic, Start: 13.5, End: 13.959999999999999
```

Figure: List of all Words identified with timestamps

Results

```
Iterating through swear word list and muting...
##########
Swear tuple: (' fucking', 4.92, 5.239999999999999)
Swear tuple: (' fuck', 14.36, 15.24)
Swear tuple: (' suck', 17.04, 17.360000000000003)
Swear tuple: ('dick', 17.5, 17.880000000000003)
##########
Muting all F-words...
```

Figure: Identified Word tuples to be censored

Results

Figure: Example of an image depicting gun violence

Results

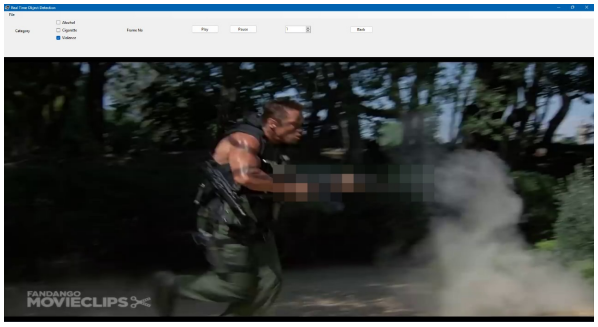

Figure: After the censorship/blurring of the gun

Results

Figure: Example of substance abuse: A man smoking a cigarette

Results

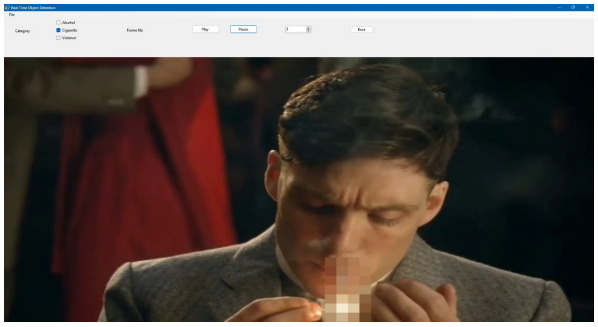

Figure: After the censorship/blurring of the cigarette

Results

```
# Example text input
input_txt="Ana,a college student,interviews an enigmatic billionaire entrepreneur,Christ

# Preprocess the input text
preprocessed_text = preprocess_text(input_text)
```

```
1/1 [==============================] - 0s 49ms/step
Predicted genres: ['Comedy', 'Romance', 'Drama']
```

Figure: Genre Identification using transcripts

Results

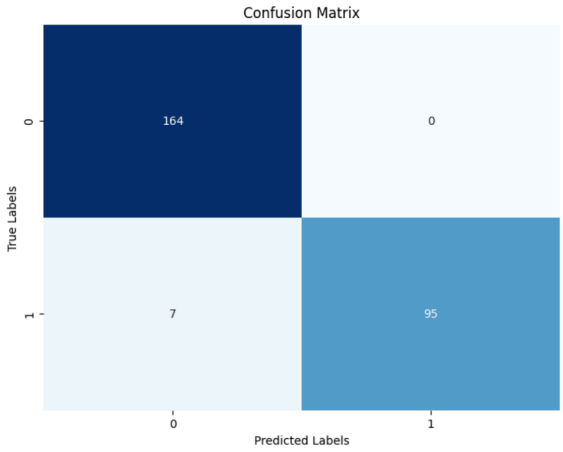

Figure: Confusion Matrix of HAR module

Results

- Accuracy of Genre Classification Module is 85.24%
- Loss value of HAR module is 0.067
- Combined accuracy of YOLOv8 on all explicit classes was found to be 92.15 percent.

Conclusion

Conclusion

- The development of a comprehensive multimodal content analysis system is essential for safeguarding viewers and users from harmful content.
- By effectively identifying and classifying inappropriate content across various modalities, it ensures a safer and more responsible digital environment.
- The proposed system leverages deep learning and natural language processing techniques to achieve its goals, providing a robust and scalable solution for content analysis.

Future Scope

Future Scope

- Integrate multilanguage support into the audio censoring system to expand its applicability.
- Acquire additional datasets to enhance the accuracy and effectiveness of the movie censoring system.
- Implement user-driven censoring functionality, allowing users to flag inappropriate content and enabling the model to automatically censor it based on user preferences.

References

References

- R. Nar, A. Singal, and P. Kumar, "Abnormal activity detection for bank atm surveillance," in 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2016, pp. 2042–2046.
- J. Xie, W. Yan, C. Mu, T. Liu, P. Li, and S. Yan, "Recognizing violent activity without decoding video streams," Optik - International Journal for Light and Electron Optics, vol. 127, 11 2015.
- A. S. Yuksel and F. G. Tan, "Deepcens: A deep learning-based system for real-time image and video censorship," Expert Systems, p. e13436, 2023.
- M. M. Taha, A. B. Zaky, and A. W. Alsammak, "Filtering of inappropriate video content a survey," International Journal of Engineering Research & Technology (IJERT), vol. 11, no. 2, 2022.

References

- S. U. Khan, I. U. Haq, S. Rho, S. W. Baik, and M. Y. Lee, "Cover the violence: A novel deep-learning-based approach towards violence-detection in movies," Applied Sciences, vol. 9, no. 22, 2019. [Online]. Available: https://www.mdpi.com/2076-3417/9/22/4963
- N. Fei and Y. Zhang, "Movie genre classification using tf-idf and svm," in Proceedings of the 2019 7th International Conference on Information Technology: IoT and Smart City, ser. ICIT '19. New York, NY, USA: Association for Computing Machinery, 2020, p. 131–136. [Online]. Available: https://doi.org/10.1145/3377170.3377234
- A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in Proceedings of the 40th International Conference on Machine Learning, ser. ICML'23. JMLR.org, 2023.

References

- D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA, USA: IEEE Computer Society, jun 2018, pp. 6450–6459. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00675
- G. Chen, S. Chai, G.-B. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang, M. Jin, S. Khudanpur, S. Watanabe, S. Zhao, W. Zou, X. Li, X. Yao, Y. Wang, Z. You, and Z. Yan, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," 08 2021, pp. 3670–3674.

Paper Submission

Paper Submission

- Submitted to International conference on Artificial Intelligence (ICAI) conducted by Advanced Research Society for Science and Sociology (ARSSS) conducted on 29 April 2024. Indexed by SCOPUS.
  - Status of application : Accepted
  - Current status : Paper has been retracted
- Submitted to 4th IEEE International Conference on Artificial Intelligence and Signal Processing-AISP '24 organized by the Vellore Institute of Technology, Andhra Pradesh on 26-28 October 2024. Indexed by SCOPUS, Web of Science, Elsevier and Springer
  - Status of application : Submitted
  - Current status : Awaiting further communication

Thank you!

Appendix B: Vision, Mission, Programme Outcomes and Course Outcomes

Vision, Mission, Programme Outcomes and Course Outcomes

Institute Vision

To evolve into a premier technological institution, moulding eminent professionals with creative minds, innovative ideas and sound practical skill, and to shape a future where technology works for the enrichment of mankind.

Institute Mission

To impart state-of-the-art knowledge to individuals in various technological disciplines and to inculcate in them a high degree of social consciousness and human values, thereby enabling them to face the challenges of life with courage and conviction.

Department Vision

To become a centre of excellence in Computer Science and Engineering, moulding professionals catering to the research and professional needs of national and international organizations.

Department Mission

To inspire and nurture students, with up-to-date knowledge in Computer Science and Engineering, ethics, team spirit, leadership abilities, innovation and creativity to come out with solutions meeting societal needs.

Programme Outcomes (PO)

Engineering Graduates will be able to:

1. Engineering Knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

2. Problem analysis: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

3. Design/development of solutions: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

4. Conduct investigations of complex problems: Use research-based knowledge including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

5. Modern Tool Usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

6. The engineer and society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal, and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

7. Environment and sustainability: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

8. Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

9. Individual and Team work: Function effectively as an individual, and as a member or leader in teams, and in multidisciplinary settings.

10. Communication: Communicate effectively with the engineering community and with society at large. Be able to comprehend and write effective reports documentation. Make effective presentations, and give and receive clear instructions.

11. Project management and finance: Demonstrate knowledge and understanding of engineering and management principles and apply these to one's own work, as a member and leader in a team. Manage projects in multidisciplinary environments.

12. Life-long learning: Recognize the need for, and have the preparation and ability to engage in independent and lifelong learning in the broadest context of technological change.

Programme Specific Outcomes (PSO)

A graduate of the Computer Science and Engineering Program will demonstrate:

PSO1: Computer Science Specific Skills
The ability to identify, analyze and design solutions for complex engineering problems in multidisciplinary areas by understanding the core principles and concepts of computer science and thereby engage in national grand challenges.

PSO2: Programming and Software Development Skills
The ability to acquire programming efficiency by designing algorithms and applying standard practices in software project development to deliver quality software products meeting the demands of the industry.

PSO3: Professional Skills
The ability to apply the fundamentals of computer science in competitive research and to develop innovative products to meet the societal needs thereby evolving as an eminent researcher and entrepreneur.

Course Outcomes (CO)

Course Outcome 1: Model and solve real world problems by applying knowledge across domains (Cognitive knowledge level: Apply).

Course Outcome 2: Develop products, processes or technologies for sustainable and socially relevant applications (Cognitive knowledge level: Apply).

Course Outcome 3: Function effectively as an individual and as a leader in diverse teams and to comprehend and execute designated tasks (Cognitive knowledge level: Apply).

Course Outcome 4: Plan and execute tasks utilizing available resources within timelines, following ethical and professional norms (Cognitive knowledge level: Apply).

Course Outcome 5: Identify technology/research gaps and propose innovative/creative solutions (Cognitive knowledge level: Analyze).

Course Outcome 6: Organize and communicate technical and scientific findings effectively in written and oral forms (Cognitive knowledge level: Apply).

Appendix C: CO-PO-PSO Mapping

CO-PO AND CO-PSO MAPPING

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12	PSO 1	PSO 2	PSO 3
CO 1	2	2	2	1	2	2	2	1	1	1	1	2	3		
CO 2	2	2	2		1	3	3	1	1		1	1		2	
CO 3									3	2	2	1			3
CO 4					2			3	2	2	3	2			3
CO 5	2	3	3	1	2							1	3		
CO 6					2			2	2	3	1	1			3

3/2/1: high/medium/low

JUSTIFICATIONS FOR CO-PO MAPPING

MAPPING	LOW/MEDIUM/ HIGH	JUSTIFICATION
100003/ CS722U.1-PO1	M	Knowledge in the area of technology for project development using various tools results in better modeling.
100003/ CS722U.1-PO2	M	Knowledge acquired in the selected area of project development can be used to identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions.

100003/ CS722U.1-PO3	M	Can use the acquired knowledge in designing solutions to complex problems.
100003/ CS722U.1-PO4	M	Can use the acquired knowledge in designing solutions to complex problems.
100003/ CS722U.1-PO5	H	Students are able to interpret, improve and redefine technical aspects for design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
100003/ CS722U.1-PO6	M	Students are able to interpret, improve and redefine technical aspects by applying contextual knowledge to assess societal, health and consequential responsibilities relevant to professional engineering practices.
100003/ CS722U.1-PO7	M	Project development based on societal and environmental context solution identification is the need for sustainable development.
100003/ CS722U.1-PO8	L	Project development should be based on professional ethics and responsibilities.
100003/ CS722U.1-PO9	L	Project development using a systematic approach based on well defined principles will result in teamwork.
100003/ CS722U.1-PO10	M	Project brings technological changes in society.

100003/ CS722U.1-PO11	H	Acquiring knowledge for project development gathers skills in design, analysis, development and implementation of algorithms.
100003/ CS722U.1-PO12	H	Knowledge for project development contributes engineering skills in computing & information gatherings.
100003/ CS722U.2-PO1	H	Knowledge acquired for project development will also include systematic planning, developing, testing and implementation in computer science solutions in various domains.
100003/ CS722U.2-PO2	H	Project design and development using a systematic approach brings knowledge in mathematics and engineering fundamentals.
100003/ CS722U.2-PO3	H	Identifying, formulating and analyzing the project results in a systematic approach.
100003/ CS722U.2-PO5	H	Systematic approach is the tip for solving complex problems in various domains.
100003/ CS722U.2-PO6	H	Systematic approach in the technical and design aspects provide valid conclusions.
100003/ CS722U.2-PO7	H	Systematic approach in the technical and design aspects demonstrate the knowledge of sustainable development.

100003/ CS722U.2-PO8	M	Identification and justification of technical aspects of project development demonstrates the need for sustainable development.
100003/ CS722U.2-PO9	H	Apply professional ethics and responsibilities in engineering practice of development.
100003/ CS722U.2-PO11	H	Systematic approach also includes effective reporting and documentation which gives clear instructions.
100003/ CS722U.2-PO12	M	Project development using a systematic approach based on well defined principles will result in better teamwork.
100003/ CS722U.3-PO9	H	Project development as a team brings the ability to engage in independent and lifelong learning.
100003/ CS722U.3-PO10	H	Identification, formulation and justification in technical aspects will be based on acquiring skills in design and development of algorithms.
100003/ CS722U.3-PO11	H	Identification, formulation and justification in technical aspects provides the betterment of life in various domains.
100003/ CS722U.3-PO12	H	Students are able to interpret, improve and redefine technical aspects with mathematics, science and engineering fundamentals for the solutions of complex problems.

100003/ CS722U.4-PO5	H	Students are able to interpret, improve and redefine technical aspects with identification formulation and analysis of complex problems.
100003/ CS722U.4-PO8	H	Students are able to interpret, improve and redefine technical aspects to meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
100003/ CS722U.4-PO9	H	Students are able to interpret, improve and redefine technical aspects for design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
100003/ CS722U.4-PO10	H	Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools for better products.
100003/ CS722U.4-PO11	M	Students are able to interpret, improve and redefine technical aspects by applying contextual knowledge to assess societal, health and consequential responsibilities relevant to professional engineering practices.
100003/ CS722U.4-PO12	H	Students are able to interpret, improve and redefine technical aspects for demonstrating the knowledge of, and need for sustainable development.
100003/ CS722U.5-PO1	H	Students are able to interpret, improve and redefine technical aspects, apply ethical principles and commit to

		professional ethics and responsibilities and norms of the engineering practice.
100003/ CS722U.5-PO2	M	Students are able to interpret, improve and redefine technical aspects, communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.
100003/ CS722U.5-PO3	H	Students are able to interpret, improve and redefine technical aspects to demonstrate knowledge and understanding of the engineering and management principle in multidisciplinary environments.
100003/ CS722U.5-PO4	H	Students are able to interpret, improve and redefine technical aspects, recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.
100003/ CS722U.5-PO5	M	Students are able to interpret, improve and redefine technical aspects in acquiring skills to design, analyze and develop algorithms and implement those using high-level programming languages.
100003/ CS722U.5-PO12	M	Students are able to interpret, improve and redefine technical aspects and contribute their engineering skills in computing and information engineering domains like network design and administration, database design and

		knowledge engineering.
100003/ CS722U.6-PO5	M	Students are able to interpret, improve and redefine technical aspects and develop strong skills in systematic planning, developing, testing, implementing and providing IT solutions for different domains which helps in the betterment of life.
100003/ CS722U.6-PO8	H	Students will be able to associate with a team as an effective team player for the development of technical projects by applying the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
100003/ CS722U.6-PO9	H	Students will be able to associate with a team as an effective team player to Identify, formulate, review research literature, and analyze complex engineering problems
100003/ CS722U.6-PO10	M	Students will be able to associate with a team as an effective team player for designing solutions to complex engineering problems and design system components.
100003/ CS722U.6-PO11	M	Students will be able to associate with a team as an effective team player, use research-based knowledge and research methods including design of experiments, analysis and interpretation of data.
100003/ CS722U.6-PO12	H	Students will be able to associate with a team as an effective team player, applying ethical principles and

		commit to professional ethics and responsibilities and norms of the engineering practice.
100003/ CS722U.1-PSO1	H	Students are able to develop Computer Science Specific Skills by modeling and solving problems.
100003/ CS722U.2-PSO2	M	Developing products, processes or technologies for sustainable and socially relevant applications can promote Programming and Software Development Skills.
100003/ CS722U.3-PSO3	H	Working in a team can result in the effective development of Professional Skills.
100003/ CS722U.4-PSO3	H	Planning and scheduling can result in the effective development of Professional Skills.
100003/ CS722U.5-PSO1	H	Students are able to develop Computer Science Specific Skills by creating innovative solutions to problems.
100003/ CS722U.6-PSO3	H	Organizing and communicating technical and scientific findings can help in the effective development of Professional Skills.