National College of Ireland

Facial Emotion and Behavioural Recognition From Video Using Deep learning approach

MScResearchProject
Data Analytics

Sugandha Sharma
StudentID:x21236577

SchoolofComputing
NationalCollegeofIreland

Supervisor: Athanasios Staikopoulos

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Sugandha Sharma
Student ID:	x21236577
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Athanasios Staikopoulos
Submission Due Date:	14/12/2023
Project Title:	Facial Emotion and Behavioural Recognition from Video Using Deep learning approach
Word Count:	5452
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	30th January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Facial Emotion and Behavioural Recognition from Video Using Deep learning approach

Sugandha Sharma
x21236577

Abstract

Security cameras have become essential now a days in public areas or high-risk areas for the security concerns. To ensure this it is important to capture the emotions and actions of the individuals to take an appropriate action in case of any unexpected events. This analysis will help to understand the behavioural characteristics of an individual to prevent crimes. In addition, face detection is implemented in case of tracking purpose. The primary objective of this research is to analyse emotions, actions, and face on video dataset and in real time scenario to understand the behavioural characteristics of individual. Several models like 3D CNN model, RNN, pretrained model like Deep Face, Haar Cascade Classifier, MTCNN to predict the emotions and actions by individuals. The study emphasizes on importance of these models in real time scenarios which is a contribution in security sector. It highlights the technical aspects and the role of technology in enhancing the security systems. The performance of classifier has improved with the large dataset. When models are applied to real-time settings as compared to pre-trained datasets it tends to yield better results.

1 Introduction

1.1 Background

Emotions and actions of individual plays a vital role in understanding the intentions of human being. This approach has been used to implement the security roles. The aim of this research is to identify emotions and actions of humans to understand the intentions and behaviour of individual in video. This process will be beneficial in security surveillance areas. In cases of threat, harass, panic mode right amount of action can be taken if the intentions are known. Further the identification of individuals in existing database is done. This will help in tracking people, management of the VIP security. Many models have been created to detect emotions in images but not in videos. In this research I have tried to detect emotions and classify actions in same video and detect the faces to match with the existing dataset. Also, the study is done on real time scenarios for in depth analysis. This research has been extended to videos for better understanding of human behaviour. Many deep learning models and pretrained models have been used to depict the emotions and actions of the individual. Face detection model will be used to identify the face from existing database. This approach will save time during the tracking cases. The motivation to implement deep learning techniques to enhance the surveillance security system led to research on this topic.

1.2 Research Question

The research question for this study is:

"How deep learning helps in extracting emotions and actions of individuals using videos and verify the identities by comparing the image in existing database to enhance security measures?"

1.3 Research Objective

The main aim of this research is to detect emotions and actions of individual in live video. This will help to understand the characteristics of any individual in depth. In case of negative scenario, it can help to detect the situation and alarm the third party for help. This solution can be utilized in criminal investigation, security management and psychological analysis. For an instance during VIP security by detecting faces one can decide who is allowed to enter inside the premises. In public areas like public transport, shopping malls this study will help to analyse if there is any tension in any situation or it is a calm environment. In case of any crime this model would assist to understand how many people were present at that scene and what kind of actions they were doing. The dataset has been taken from an opensource known as "BOSS dataset[1]. It contains different videos from local train. Deep learning models and pretrained models are used to predict the results and compare with the real time scenarios.

1.4 Structure of Document

The section I is Introduction that gives the overview of the research topic. It tells the research question and objective to explain the reason behind this research. In section III related work written by different authors have been discussed and tells how their work has motivated and guided to implement the models to continue the research. It also contains the critical analysis of their work. Section III tells about the methodology that have been used to conclude this research. It gives the glimpses of how data is selected. Pre-processing datasets, models used, evaluation steps. Section IV tells the design specification. It discusses the techniques; tools used during the thesis and gives brief overview of the architecture that is used. Section V is the implementation section where I have discussed transformed data, code. It discusses the final stage of the implementation. Section VI shows the results obtained after evaluating the models. Different model's results have been compared for performing critical analysis on them and infer which model has the best accuracy. Section VII tells the conclusion and future work. In conclusion, findings related to the research is mentioned and explained how successful it is to answer the research question. In future work I have discussed the limitations of the project and suggested what solutions can be proposed to remove the drawbacks.

[1] http://velastin.dynu.com/videodatasets/BOSSdata/whole_dataset.html#Faces_1.

2 Related Work

The academic research is mainly focused on detecting emotions, classifying different action and detecting the faces from video dataset. It also focus on methodologies and techniques mentioned in previous papers.

2.1 Techniques to classify the actions.

The effectiveness of 3D CNNs in the residual learning framework and questions the current dependence on 2D CNNs for video action detection. The spatiotemporal convolutional block is introduced in the research(Tran, D., 2018) . It highlights the merging of 3D and 2D convolutions for increased accuracy by mixed convolution and (2+1)D convolution block.

The research paper (B.Patel, 2023) explains the incorporation of cameras into a wide range of social contexts, including intelligent spaces, parking lots, live traffic monitoring, and space surveillance, has grown in popularity. This study talks about the recurrent artificial neural networks to examine human behaviour in video clips and categorize it as safe or dangerous. The suggested method seeks to automate the evaluation process by locating pertinent elements and forecasting possible actions through- out video sequences. Real-time human behaviour recognition improves surveillance capabilities by streamlining the analysis of routine activities recorded by cameras and enabling quick decisions and actions based on the recognized behaviours. Although this paper covers the human behaviour in real time however it limits to the emotions of the individual.

This paper (LokeshNaik, 2023)discusses about the deep learning temporal relation module to extract motion information through the entire video.2D CNN has been used to increase the accuracy of the action.

This study (Morshed, M.G et al.; 2023) present human activity recognition in computer vision using RGB and depth data. Their focus is to recognize human action in real world.

This study (Yang,Xiao et al., 2019) is on raw depth video projected with different virtual imaging viewpoints by rotating virtual camera in 3D space.

This paper (Yao Guangle et al.,2019) has used CNN to extract 2D spatial features from still image and videos as 3D spatiotemporal signals

2.2 Papers discussing on Face Detection

The research study (G. Barquero, 2021) highlights the architecture for face tracking. This study highlights the shortcomings of past approaches and strive to provide reliable and accurate face tracking in difficult scenarios.

Another research paper (J.M.Llaurado´, 2023) discussed about the problems occur in face recognition during the surveillance and security in smart city. It highlights the advancement in facial recognition technology and emphasized on the effects of low-resolution photos, sensor quality and ethical issues that occur while implementing facial recognition in smart cities.

2.3 Papers discussing on Facial Emotion Recognition

The literature study (Nemade & Gohokar, 2017) delves deeply into the field of facial expression identification, emphasizing its importance and range of uses in different fields. The basis for investigating more complicated emotional states is laid by the identification of six fundamental emotions and the division of these feelings into positive and negative categories. The author has provided a thorough analysis of the state of facial emotion identification today and points out important directions for further research.

The research paper (Kuchibhotla, 2022) discusses about Convolutional neural networks (CNNs) and facial emotion recognition (FER) which are important fields of study with the goal of classifying human emotional states from facial expressions. The use of CNNs to categorize facial images into seven different emotion classes—Surprise, Sad, Neutral, Happy, Fear, Disgust, and Angry—is examined in this paper. The work uses a CNN architecture with convolution and pooling layers for feature extraction, and a SoftMax layer for emotion classification. It does this by using the fer2013 dataset which is a collection of different images and real-time movies as input. The model integrates L2 regularization, dropout, and cluster standardization to handle overfitting, and it shows better accuracy than previous approaches.

This research paper (Xia, 2022) is based on developing field called Video Emotion Recognition (VER) uses text, visual, and audio modalities to infer human emotional states. This highlights the difficulty of combining global semantic information from text with the complex temporal dynamics in voice and facial emotions. The review presents a unique paradigm for emotion recognition from face videos, which departs from the conventional fusion-centric approaches. This paradigm is a multimodal interaction augmented representation learning framework. This method aligns with cognitive science findings that emotion expression is implicitly regulated by high-level semantics by incorporating a semantic enhancement module to provide text-derived semantic information to audio/visual encoders. Moreover, via cross-modal dynamic interactions, the framework uses a multimodal bottleneck Transformer to strengthen audio and visual representations.

Another study (M. A. H. Akhand, 2021) presents on audio/visual encoders implicit high level of emotion expression and text-derived semantic information. It integrates text, visual, and audio modalities to identify human emotional states. Previous methods frequently struggle to reconcile the complex temporal dynamics present in voice and facial emotions with the global semantic information from text.

In particular, the dynamic information found in video sequences is the main emphasis of this article's (M. Nasir, 2023) tells contextual analysis of nonverbal communication, which delves into the complex field of human emotion identification.
This research (M. Nasir, 2023) interprets human emotion by generating geometric features based on prominent landmarks on face photos, the suggested automatic identification method presents a triangulation mechanism. The active appearance model (AAM) is utilized by the system to monitor landmark points in a series of image frames. Four centre points are identified from triangles produced by specific face points. Different geometric representations of emotional image sequences are signified by six unique geometric distance signatures that are generated. To identify six fundamental emotions, a multilayer perceptron (MLP) classifier uses each of these distance signatures separately as input feature sets.

A crucial component of video surveillance and public safety management is crowd behaviour analysis; however, recent developments have mostly concentrated on event detection rather than population emotional state assessment. To bridge this gap, this research (X. Zhang, 2021) suggests a technique for efficiently assessing crowd emotion that makes use of fuzzy inference based on the arousal-valence model. To accommodate uncertainty in crowd attributes, the suggested approach integrates features such as confusion index, enthalpy, magnitude variance, and crowd density into a fuzzy inference system to represent crowd emotion. The method's effectiveness in evaluating arousal and valence is supported by experimental results, which provide important insights into the emotional dynamics of crowds in situations involving video monitoring. This paper insights help me in predicting the actions of the crowd.

This study focus (Ullah, et al., 2017) on different techniques like CNN to recognize seven fundamental emotions. This model is efficient in terms of emotional recognition.

2.4 Critical Analysis of Literature

The section I focuses on interpreting the actions from video however it lags to classify the emotions from them. The other section talks about detecting the face for security purposes but it will not interpret the emotions and actions of the individual. In the last section although emotions have been recognized by using different methodologies however still it does not cover all the three scenarios together.

In this paper I am analysing the emotions and actions of humans from video datasets and real time scenarios to have better understanding of the characteristics of an individual. Also identifying the images from dataset. This study will help to improve the security systems and can be utilized in different sectors like retail, public.

3 Methodology

This section gives an overview of methodologies that are required to achieve the results. The main objective is to recognize the emotions, classify the actions and detect the faces in video.

3.1 Dataset

The dataset for this research has been taken from the BOSS dataset. It is an open source dataset. This video dataset contains the video of 10 to 12 people travelling in a moving train. The videos have been captured through the security cameras in the train. This is basically captured to understand the different activities happen in train and to reduce the crime rates. This video contains different actions like sitting, walking, laying down, fighting.

3.2 Data Cleaning and Pre-processing

The videos are in .avi format which is converted into .mp4 formats. The videos are manually distributed into different folders Fighting, Walking, Sitting and Laying Down. Further the videos are extracted into frames. These frames are saved in different folder named as “output faces”. This dataset is divided into train and validation sets. These are total 4424 files out of which 3540 are for training set ,884 for validation set. Then model is trained on the training set. Adam optimizer is used during compilation and is trained for 30 epochs. Further the training and validation loss and accuracy graphs are generated. The data is augmented to increase size of training dataset irrespective of the labels. The augmented images are shown in Figure 1

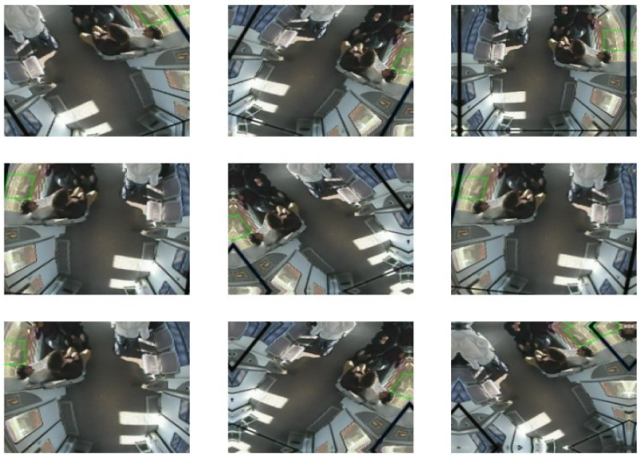

Figure 1: Augmented Images

The model is trained, and graphs are generated from training and validation accuracy.

Three different directories are created in a subset test, train, Val for labelled dataset. The videos are split into these three directories. Dataframes are created and two csv files are created (test and train) with column names as tag and video name. tags column are the labels. Further new column is added into the dataframes named as ‘label ‘where Fighting': 0, 'Laying Down': 1,'Sitting':2,'Walking':3. Then features are extracted using keras and labels are processed. Hyperparameters are defined (image size, epochs, batch size, num_features). Frame features and frame masks are extracted from train set. Further model is trained for 100

epochs. When the video is tested it gives results sitting, Laying Down, Walking, Fighting. In Table 1 values are given.

Table 1: Action Classification

Video	**Value (%)**
Sitting	25.01
Walking	25.01
LayingDown	25.01
Fighting	24.96

The graph of the Training and Validation Accuracy and Loss are shown in Figure 2

Figure 2: Graph Diagram

Video classification is done using 3D convolutional neural network. In this video are loaded and pre-processed. Frame Generator class is used to load video frames with the encoded labels. After that 3D model is created. Convolution layers are applied to spatial dimension and later temporal dimension. ResNet model is used to add input

to output of the main branch. Normalization layer is added and activation function ReLU is used. The video is resized to down sample the data. Further Keras functional API is used to build the residual network. Adam optimizer is used for this model, and it is trained on training dataset. The results accuracy is visualized as a graph. At the end test data is evaluated by analysing confusion matrix.
Face Detection and emotion Detection on live webcam is done by Deep Face and CV2 library. Pretrained emotion model is used. Haar cascade classifier is used for detecting the face. The image is reshaped for model input. Rectangle is drawn around face, emotions are predicted using OpenCV functions.

3.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a crucial step in the data analysis process. It involves the use of statistical and visual methods to examine, summarize, and understand the main characteristics of a dataset. The primary goals of EDA are to uncover patterns, identify trends, detect outliers, and gain insights into the underlying structure of the data. It aims to summarize the patterns, trends, and outliers in the structure of data. It assesses the data quality and finds out the missing values.

3.4 Model Evaluation

Classification model is used to evaluate the results. The common metrics used for this accuracy, precision, F1 score, recall and confusion matrix.

4 Design Specification

This research is done on BOSS dataset where videos are pre-processed, and frames are generated. Further these frames are analysed and split into train, test, and validation datasets. Different CNN and pretrained models are applied on them to detect the emotions and classify the actions in the videos. Along with the dataset I have used one recorded video and real time webcam scenarios to obtain the results from real time situation as well. These findings help to understand how the safety process will be strengthen in risk prone areas. The models are compiled and run epochs on training dataset. At the end the results of emotion detection, action classification and face detection will be compared to verify which scenario gives better results to understand the model performance.

4.1 Model Functionality

- Open Face: Pretrained Model that help to detect faces and also its libraries help to detect the emotions in a live view. It provides features for emotion analysis as well.
- 3D CNN model : It is based on the CNN. Multiple visual frames are fed into a standard 2D CNN, which processes and analyses them. When spatial, temporal, and spatial discretization are included in the input of a 3D convolutional neural network, it is feasible to examine the relationship between frames in time. From

the input, the convolutional layers identify and extract temporal and spatial characteristics. By adding non-linearity to the model, activation functions enable it to represent intricate interactions between the input and the desired output. An input layer, numerous convolutional layers, activation functions, max-pooling layers, and a final classification layer make up the architecture of a 3D convolutional neural network

- The CNN model (Chua, L.O.,1998)is implemented using Keras Sequential API which is used for image classification. So it is implemented on output_faces. Rescaling Layer : The input layer images are divided by 255 to rescale the photos.
  Convolutional Layers has 16 filters and ReLU activation. Max pooling layer reduce the spatial dimension of feature maps.
  Flatten Layers flattens fully connected layers from 3D to 1D vector.
  Dense Layers are the output layer with a neuron count equal to the number of classes. The several convolutional layers in the model are for feature extraction, max-pooling layer are for down-sampling and fully connected layers are for classification. ReLU activation introduces non-linearity and final prediction is determined by SoftMax activation during training.
- RNN model : The code provided is a Recurrent Neural Network (RNN) model with Gated Recurrent Unit (GRU) layers, specifically intended for processing sequence data. It accepts sequences of frame features as input and handles variable-length sequences by using a mask. A dropout layer aids in preventing overfitting, while the GRU layers capture temporal connections within the sequences. Through a series of dense layers, the model reduces the sequence to a final output vector, which is then used for multi-class classification through a SoftMax activation. This design is appropriate for sequential data problems, such video classification, where precise prediction depends on comprehending the temporal links between frames. The Adam optimizer and categorical cross-entropy loss are used to train the model.
- Haar Cascade Classifier – It is a face detection algorithm used to detect faces in a simpler and efficient manner.
- MTCNN(Multi -task Cascaded Convolutional Network) – It is a three layer network that refines the face candidate. It is basically used for real time detection. It is preferred for accuracy.
  The CNN architecture is shown in Figure 3.

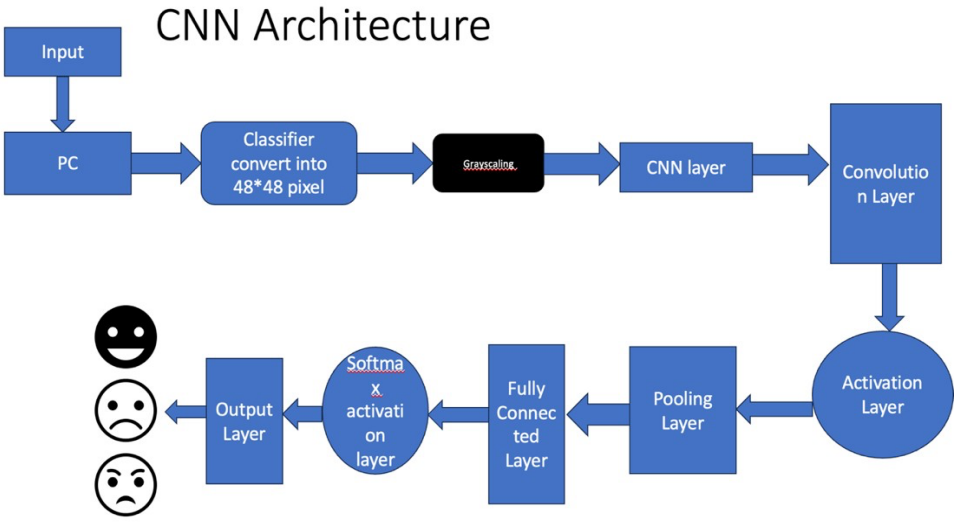

Figure 3: CNN Architecture

This explains how the input is passed to computer with web camera. Then that image is converted into 48*48 pixel. Further it is grayscale and passed to CNN, Convolution layer, Activation layer, pooling layer, fully connected layer, SoftMax activation layer , output layer and finally detects the emotions.

4.2 Evaluation Techniques

- Accuracy : It is a metrics to evaluate the performance of model. It is number of correct predictions by total number of predictions.
- Confusion Matrix: It summarizes the performance of the model and breakdown the predictions and actual outcomes. It consist of True Positive, True Negative, False Positive and False Negative.
- Precision and Recall: Precision measure the accuracy of positive results. Recall means True Positive Rate. These are basically used for imbalance dataset.

5 Implementation

5.1 Environment Set Up

The research is implemented on Google Collab. TensorFlow[1] and Keras[2] are used for deep learning models. Scikit-learn, Matplotlib[3] is used for visualizations. The dataset is BOSS dataset taken from open source. OpenCV is required for video processing tqdm for process bar. BOSS dataset is used for processing videos.

5.2 Data Handling

Datasets are examined and properly analysed. Early stopping is used for avoiding overfitting if validation loss is unchanged. All the videos were not included. Empty frames[4] were removed from the dataset. Only the important videos with labels are included. For image classification the

[1] https://www.tensorflow.org/api docs/python/tf/keras/layers

[2] https://keras.io/api/applications/

3 https://matplotlib.org/stable/plot types/index

4 https://www.hackersrealm.net/post/facial-emotion-recognition-using-python

dataframes were taken from few videos and segregated accordingly. For face detection any random video is taken from the datasets and then it is pre-processed to detect face. For the better performance of the model hyperparameters are optimized. It increases the accuracy.

5.3 Implementing Models

• Identification of emotions— The video is divided into different frames. Extracted5 frames are saved as images in one folder. Further pretrained model is applied on the image to detect the emotions in that image (Sun, Y., Sebe,2004). The count of Actions in dataset is shown in Figure 3. This tells the count of different action videos in the datasets. Walking has the least number of videos.

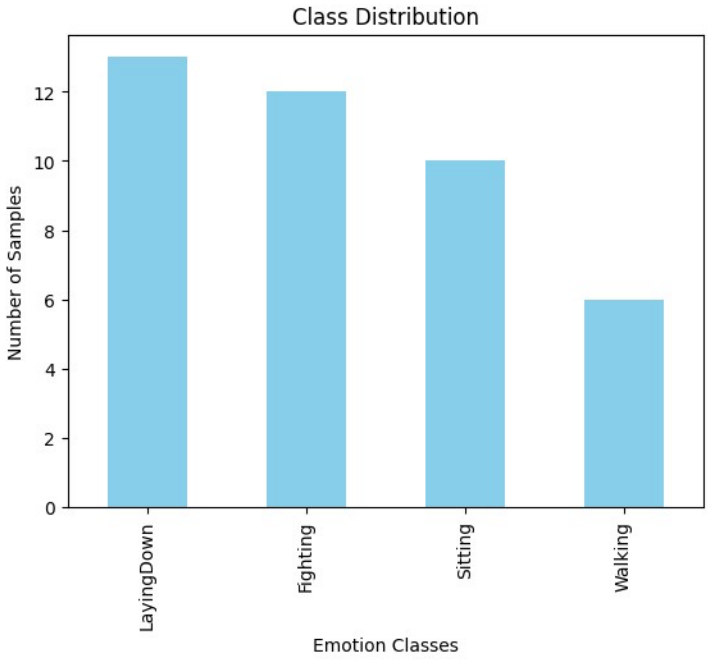

Figure 4: Distribution of Actions

- **Classification of Action:** Firstly, sequential model is used without labels. 3D CNN is used to classify the actions. Input pipeline is built. Then data is pre-processed, and frames are extracted, resized, normalized, augmented. Model is Assembled using optimizer and loss function. Model is trained using predefined dataset to avoid overfitting. At the end, model is analysed using test dataset.

- **Face Detection:** For face detection compared two different images by applying the pretrained model if the person is same or different. This will help in face detection. In times of any crime, this will easily help to find out the victim and attacker. Due to the wide variety of human faces, face identification algorithms usually require extensive input data training to achieve accuracy. After importing the package, images are read and converted to grayscale. Finally, the classifier is loaded, and face detection is performed.

- **Real time Camera detection** – This study is for the real time scenarios. Hence experimented with the web camera. In this OpenCV and the `deep face` package is used to create real-time facial emotion recognition. The aim is to record video in real time using a camera, recognize faces in the feed, and forecast the emotions for every face that is identified. On the video frames, the anticipated emotions are shown in real time. Important files are loaded then Haar Cascade classifier (Hashim, S. et al.; 2023) load XML file and process the loop. Frames are transformed to grayscale to detect image.
  Finally, the images are pre-processed, and emotion is detected.

5.4 Exploratory Data Analysis

This is the Emotion probabilities Over time in Figure 5. The different colours depict different emotions passed over the time.

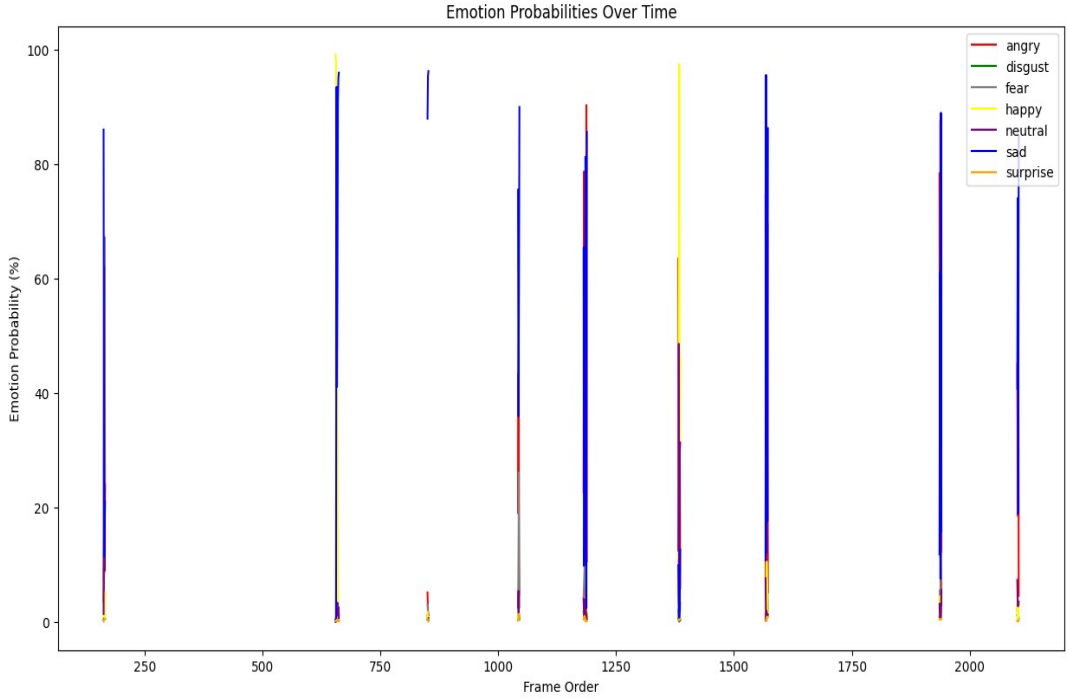

Figure 5: Emotional Probabilities Over Time

This is the Video Duration Distribution in Figure 6. It gives the count of videos over different durations. Most number of videos are for 100 seconds.

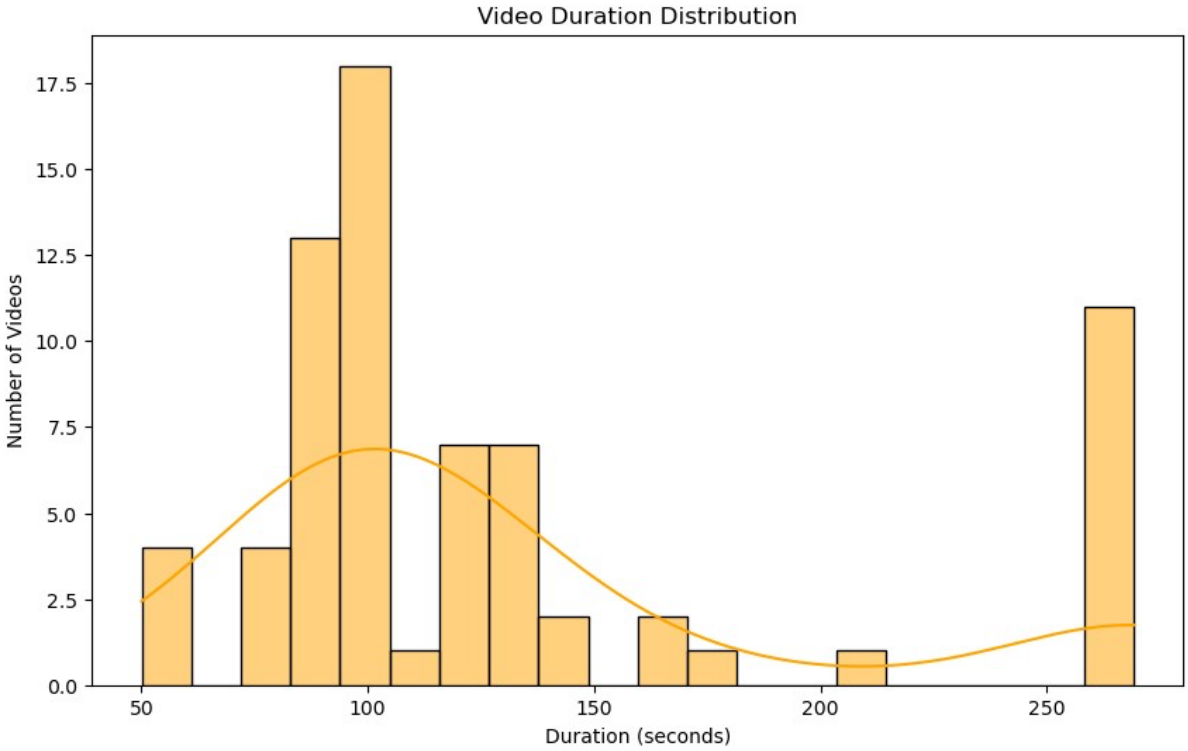

Figure 6: Video Duration Distribution

6 Evaluation

The purpose of this section is to evaluate the results based on different experiments.

The purpose of this section is to provide a comprehensive analysis of the results and main findings of the study as well as the implications of these finding both from academic and practitioner perspective are presented. Only the most relevant results that support your research question and objectives shall be presented. Provide an in-depth and rigorous analysis of the results. Statistical tools should be used to critically evaluate and assess the experimental research outputs and levels of significance.

Use visual aids such as graphs, charts, plots and so on to show the results.

6.1 Experiment 1: Emotion Detection on Video Dataset and Webcam

- Video Dataset: The analysis is given in Figure 7. Although the video is blur due to moving train and different lightings but still able to confine some results.

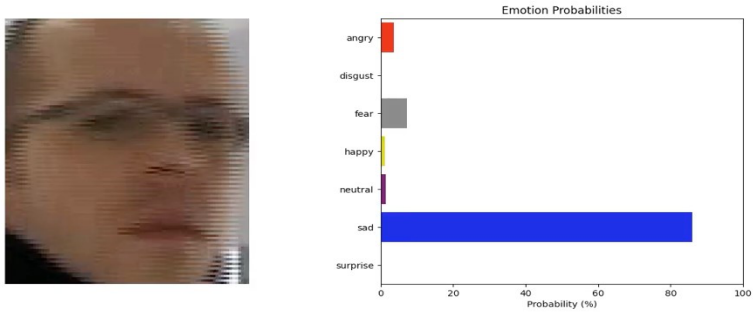

Figure 7: Emotional Probabilities Over Time in Dataset

- Webcam The analysis is given in Figure 8 and Figure 9. This depicts two emotions angry and happy.

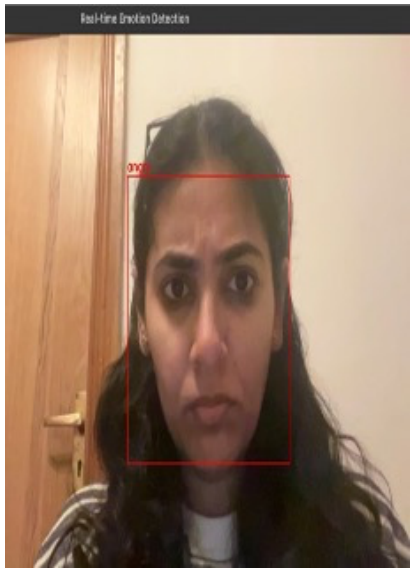

Figure 8: Angry Emotion

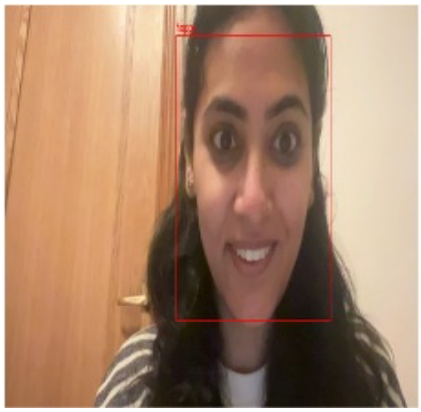

Figure 9: Happy Emotion

6.2 Experiment 2: Action Detection on Video Dataset and Webcam

- Video Dataset: The evaluation metrics of dataset in Table 2 is given below:

Metrics	**Value**
Accuracy	28.57(%)
Precision	0.28
Recall	0.285
F1 Score	0.279
Confusion Matrix	[[2 2 2 3] [1 3 1 1] [1 1 2 2] [2 2 2 1]]

Table 2: Evaluation Metrics

- Webcam: The analysis is given in Figure 10. This depicts the action of "Fixing Hair".

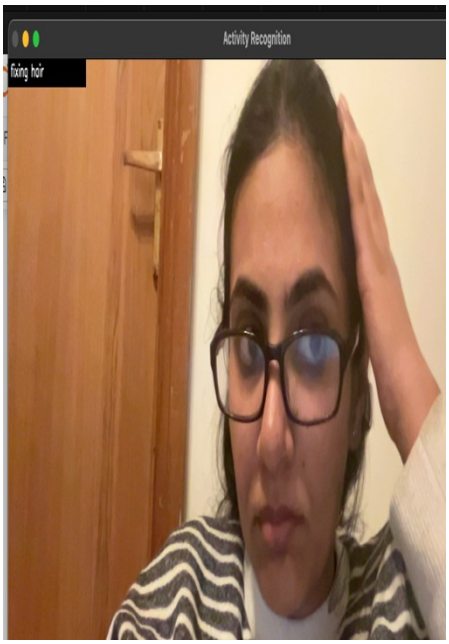

Figure 10: Action Dataset Over Time in Dataset

6.3 Experiment 3: Face Detection on Webcam

- Webcam: The analysis is given in Figure 11. This model is for face detection. It matches the existing image "Sugandha.png". As both are same it is showing as

"Known Person". If it does not match it will display "Unknown Face"

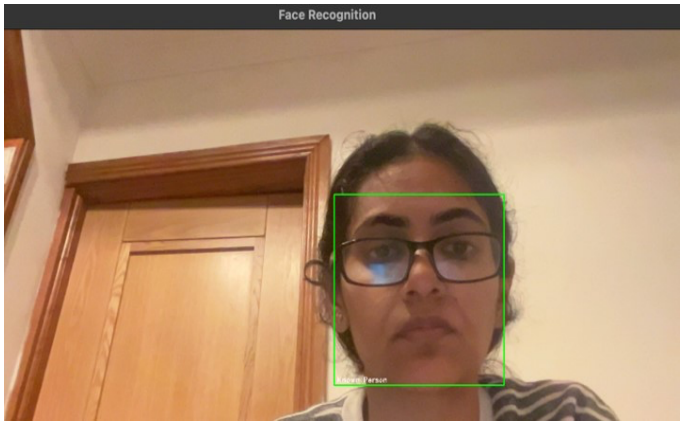

Figure 11: Face Detection in Dataset

6.4 Discussion

During the research deep learning models and few pretrained models are used to have deeper understanding of emotion and action recognition. It is observed that results obtained during real webcam give better results for action, emotion recognition and face detection. For emotion classification in datasets the result accuracy is not that good . the average quality of video recordings introduces problems like blurriness that impair the classification of actions. These findings highlight the importance of real-world situations as well as the challenges in precisely categorizing emotions within such circumstances. It becomes essential to enhance the dataset and investigate other datasets that can provide a more representative and diverse sample in order to increase classification accuracy. Classification accuracy can be increased by refining the data and explore some other datasets.

7 Conclusion and Future Work

To conclude, pre-trained models used in real time scenarios has given better results for emotion analysis, action recognition and face detection. However, difficulties occur when these models are trained on video datasets, particularly when it comes to action recognition. This can occur as there are multiple actions in one video. The data needs to be more refined and explored to increase the classification accuracy. Also, the video has many blur frames which have impacted the data quality and resulted into poor results.

The limitation of this research is it constrained in determining the multiple actions in single video. Integrating action and emotional results in real-time scenarios could complicate interpretation and necessitate additional research to determine its viability and efficacy.

In future research, the strategies can be explored for refining video datasets in case of multiple actions in a single video. Transfer learning (Sharma, P.,2021). can be focused for increasing

model performance in future. Also, can try to combine action results and emotion results in one frame during real-time scenarios. This will help to maintain the accuracy of the results and the enhancement of the model performance.

8 Acknowledgement

I am thankful to Professor Athanasios Staikopoulos for his guidance and support during this research. I also want to thank all the members of NCI for providing the platform to conduct the research. I am grateful to all of my family members who have motivated and encouraged me during the process.

References

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y. and Paluri, M. (2018). *A Closer Look at Spatiotemporal Convolutions for Action Recognition*. [online] IEEE Xplore. doi:https://doi.org/10.1109/CVPR.2018.00675.

Zhu, H., Chen, H. and Brown, R. (2018). A sequence-to-sequence model-based deep learning approach for recognizing activity of daily living for senior care. *Journal of Biomedical Informatics*, 84, pp.148–158. doi:https://doi.org/10.1016/j.jbi.2018.07.006.

ieeexplore.ieee.org. (n.d.). *IEEE Xplore Full-Text PDF:* [online] Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9647873

Llauradó, J.M., Pujol, F.A., Tomás, D., Visvizi, A. and Mar Pujol López (2023). Study of image sensors for enhanced face recognition at a distance in the Smart City context. *Scientific Reports*, 13(1). doi:https://doi.org/10.1038/s41598-023-40110-y.

ieeexplore.ieee.org. (n.d.). *A survey of video datasets for crowd density estimation | IEEE Conference Publication | IEEE Xplore*. [online] Available at: https://ieeexplore.ieee.org/document/7955333.

Kuchibhotla, S., Voonna, C., Penki, R.P., Pappala, G.S.K., S, A. and Vankayalapati, H.D. (2022). *Analysis of Facial Emotion Recognition for Image and Video Data using Convolution Neural Networks*. [online] IEEE Xplore. doi:https://doi.org/10.1109/ICECA55336.2022.10009281.

Xia, X., Zhao, Y. and Jiang, D. (2022). Multimodal interaction enhanced representation learning for video emotion recognition. 16. doi:https://doi.org/10.3389/fnins.2022.1086380.

Akhand, M.A.H., Roy, S., Siddique, N., Kamal, M.A.S. and Shimamura, T. (2021). Facial Emotion Recognition Using Transfer Learning in the Deep CNN. *Electronics*, 10(9), p.1036. doi:https://doi.org/10.3390/electronics10091036.

Nasir, M., Dutta, P. and Nandi, A. (2023). Recognition of human emotion transition from video sequence using triangulation induced various centre pairs distance signatures. *Applied Soft Computing*, [online] 134, p.109971. doi:https://doi.org/10.1016/j.asoc.2022.109971.

Zhang, X., Yang, X., Zhang, W., Li, G. and Yu, H. (2021). Crowd emotion evaluation based on fuzzy inference of arousal and valence. *Neurocomputing*, 445, pp.194–205. doi:https://doi.org/10.1016/j.neucom.2021.02.047.

Mehendale, N. (2020). Facial emotion recognition using convolutional neural networks (FERC). *SN Applied Sciences*, 2(3). doi:https://doi.org/10.1007/s42452-020-2234-1.

ieeexplore.ieee.org. (n.d.). *Emotion Recognition from Facial Expression using CNN | IEEE Conference Publication | IEEE Xplore*. [online] Available at: https://ieeexplore.ieee.org/document/9641578

ieeexplore.ieee.org. (n.d.). *Emotion Recognition from Facial Expression using CNN | IEEE Conference Publication | IEEE Xplore*. [online] Available at: https://ieeexplore.ieee.org/document/9641578

ieeexplore.ieee.org. (n.d.). *Real Time Facial Emotion Recognition using Deep Learning and CNN | IEEE Conference Publication | IEEE Xplore*.

Yi, W., Ma, S., Zhang, H. and Ma, B. (2022). Classification and improvement of multi label image based on vgg16 network, 2022 3rd International Conference on Information Science, Parallel and Distributed Systems (ISPDS), pp. 243–246.

Chua, L.O., 1998. *CNN: A paradigm for complexity* (Vol. 31). World Scientific.

Sun, Y., Sebe, N., Lew, M.S. and Gevers, T., 2004. Authentic emotion detection in real-time video. In *Computer Vision in Human-Computer Interaction: ECCV 2004 Workshop on HCI, Prague, Czech Republic, May 16, 2004. Proceedings* (pp. 94-104). Springer Berlin Heidelberg.

Sharma, P. (2021). *Transfer Learning | Understanding Transfer Learning for Deep Learning*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/10/understanding-transfer-learning-for-deeplearning/.

Hashim, S. and Mccullagh, P., 2023. Face detection by using Haar Cascade Classifier. *Wasit Journal of Computer and Mathematics Science*, *2*(1), pp.1-8.

Morshed, M.G., Sultana, T., Alam, A. and Lee, Y.K., 2023. Human Action Recognition: A Taxonomy-Based Survey, Updates, and Opportunities. *Sensors*, *23*(4), p.2182.

Xiao, Y., Chen, J., Wang, Y., Cao, Z., Zhou, J.T. and Bai, X., 2019. Action recognition for depth video using multi-view dynamic images. *Information Sciences*, *480*, pp.287-304.

Yao, G., Lei, T. and Zhong, J., 2019. A review of convolutional-neural-network-based action recognition. *Pattern Recognition Letters*, *118*, pp.14-22.

Liu, K., Liu, W., Gan, C., Tan, M. and Ma, H., 2018, April. T-C3D: Temporal convolutional 3D network for real-time action recognition. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).

Suja, P. and Tripathi, S., 2016, February. Real-time emotion recognition from facial images using Raspberry Pi II. In *2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN)* (pp. 666-670). IEEE.