JGi JAIN DEEMED-TO-BE UNIVERSITY | FACULTY OF ENGINEERING AND TECHNOLOGY

Department of Computer Science and Engineering

A Project Report on

"ENABLING SMART MACHINES TO INTERPRET HUMAN BODY LANGUAGE AND FACIAL EXPRESSION"

Submitted in partial fulfilment for the award of the degree of

MASTER OF TECHNOLOGY

IN

ARTIFICIAL INTELLIGENCE

Submitted by

SUPARNA C

(USN: 22MTRAI002)

Under the guidance of

Dr. A. RAJESH

(Professor-CSE)

Department of Computer Science and Engineering

School of Computer Science and Engineering

Faculty of Engineering and Technology

JAIN (Deemed-to-he University)

(Batch: 2022-2024)

Department of Computer Science and Engineering

A Project Report on

"ENABLING SMART MACHINES TO INTERPRET HUMAN BODY LANGUAGE AND FACIAL EXPRESSION"

Submitted in partial fulfilment for the award of the degree of

MASTER OF TECHNOLOGY

IN

ARTIFICIAL INTELLIGENCE

Submitted by

SUPARNA C

(USN: 22MTRAI002)

Under the guidance of

Dr. A. RAJESH

(Professor-CSE)

Department of Computer Science and Engineering

School of Computer Science and Engineering

Faculty of Engineering and Technology

JAIN (Deemed-to-he University)

(Batch: 2022-2024)

FACULTY OF ENGINEERING AND TECHNOLOGY

Department of Computer Science and Engineering

CERTIFICATE

This is to certify that the dissertation entitled **"ENABLING SMART MACHINES TO INTERPRET HUMAN BODY LANGUAGE AND FACIAL EXPRESSION"** is carried out by **SUPARNA C (USN: 22MTRAI002),** a bonafide student of Master of Technology, at the **School of Computer Science and Engineering, JAIN** (Deemed-to-be University) in partial fulfilment for the award of the degree of **Master of Technology in ARTIFICIAL INTELLIGENCE,** during the year **2022-2024.**

Dr. A. RAJESH,
Professor,
Department of Computer Science and Engineering,
Faculty of Engineering and Technology,
JAIN (Deemed-to-Be University)

Dr. RAJAPRAVEEN,
Associate Professor,
Department of Computer Science and Engineering,
Faculty of Engineering and Technology,
JAIN (Deemed-to-Be University)

Dr. G GEETHA,
Professor & Director,
School of Computer Science and Engineering,
Faculty of Engineering and Technology,
JAIN (Deemed-to-Be University)

Name of the examiner: signature of examiner

1.

2.

DECLARATION

I, **Ms. SUPARNA C**, a student of fourth semester M. Tech in **Artificial Intelligence** at **School of Computer Science and Engineering, JAIN** (Deemed-to-be University) hereby declare that the dissertation entitled **"Enabling smart machines to interpret human body language and facial expression"** has been carried out by me and submitted in partial fulfilment for the award of the degree of **Master of Technology in Artificial Intelligence**, during the academic year 2022-2024. Further, the matter embodied in the dissertation has not been submitted previously by anybody for the award of any degree or diploma to any University, to the best of my knowledge and faith.

PLACE: FET, JAIN GLOBAL CAMPUS SUPARNA C

DATE: USN: 22MTRAI002

ACKNOWLEDGMENT

It is a great pleasure for us to acknowledge the assistance and support of a large number of individuals who have been responsible for the successful completion of this dissertation work.

First of all, I thank Almighty God for all his blessings and thanks to my parents for giving encouragement, enthusiasm and invaluable assistance to me.

Second, I take this opportunity to express my sincere gratitude to Faculty of Engineering & Technology, JAIN (Deemed-to-be University) for providing me with a great opportunity to pursue my master's degree in this institution.

*In particular, I would like to thank **Dr. Hari Prasad S. A**., Director, Faculty of Engineering & Technology, JAIN (Deemed-to-be University) for his constant encouragement and expert advice. I am deeply grateful to him for his unwavering super guidance in every step of my dissertation work, which paved the way for smooth progress and fruitful culmination of the project.*

*I would like to thank **Dr. G Geetha**, Director, School of Computer Science and Engineering, JAIN (Deemed-to-be University) for her constant encouragement and support.*

*I would like to thank my guide **Dr. A. Rajesh**, Professor, Department of Computer Science and Engineering, School of Computer Science and Engineering, JAIN (Deemed-to-be University) for his support and encouragement during this entire tenure of my dissertation. I would also like to thank the lecturers and all the staff members of the Computer Science Department for their support and contributions.*

*It is a matter of immense pleasure to express my sincere thanks and gratitude to my guide **Dr. Raja Praveen**, Associate Professor & Program Head-PG, Department of Computer Science and Engineering, School of Computer Science and Engineering, JAIN (Deemed-to- he University) for providing right academic guidance and sparing his valuable time to extend help in every step of my dissertation work, which paved the way for smooth progress and fruitful culmination of the project.*

I would like to thank one and all who directly or indirectly helped me in completing the Dissertation work successfully.

Suparna C

USN: 22MTRAI002

ABSTRACT

This project focuses on the development of an integrated facial emotion and action recognition system that combines the capabilities of deep learning models to interpret human social cues and behaviors from video sequences. The system merges pre-trained models for facial emotion recognition and action recognition, leveraging the power of Convolutional Neural Networks (CNNs) to analyze facial expressions and recognize human actions simultaneously.

The aim of the project is to create a unified system that can accurately classify both facial emotions and physical behaviors, providing a holistic understanding of human interactions for intelligent systems. By processing video frames, predicting emotions and actions in real-time, and displaying the results on the video stream, the integrated system aims to enhance human-machine interaction and response capabilities.

Through a structured workflow involving model loading, frame preprocessing, emotion and action prediction, real-time display, user interaction, and evaluation, the project seeks to evaluate the performance of the integrated system and identify areas for improvement and future development. The system utilizes pre-trained models trained on the FER2013 dataset for facial emotion recognition and the UCF101 dataset for action recognition, demonstrating the versatility and effectiveness of deep learning in interpreting human behaviors.

The integrated facial emotion and action recognition system represents a significant advancement in human-machine interaction, offering a comprehensive solution for understanding and responding to human social cues and behaviors in real-time applications. The project aims to contribute to the field of artificial intelligence by enhancing the interpretive capabilities of intelligent systems and enabling more natural and intuitive interactions between humans and machines.

TABEL OF CONTENT

CONTENT **PAGE NO.**

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6.

CHAPTER 7

CHAPTER 8

LIST OF FIGURES

CHAPTER 1

1. INTRODUCTION

Human interaction goes beyond mere verbal communication. It encompasses a complex web of non-verbal signals such as facial expressions, body movements, and gestures. These subtle signals convey a wealth of information about emotions, intentions, and attitudes. Deciphering them is crucial for effective interaction and understanding.

The field of Artificial Intelligence (AI) have been seeing remarkable progresses in creating smart machineries! However, a major hurdle remains truly comprehending the intricate social cues that humans rely on. Existing approaches often focus solely on facial expression recognition for emotion detection or action recognition in isolation. This research aims to bridge this gap by developing an advanced system that can analyze both facial expressions and human actions simultaneously.

Utilizing advanced computer vision and deep learning methods, this project aims to close the gap between human communication and machine understanding. This will pave the way for more seamless collaboration between humans and intelligent systems.

1.2. The Landscape of Human-Computer Interaction

In today's data-driven world, human-computer interaction is flourishing. The goal is to achieve ever-increasing levels of accuracy in various tasks. With the abundance of data available, researchers are constantly striving to develop models that can extract valuable insights from specific domains. For instance, during the COVID-19 pandemic, models were created to detect mask usage in public spaces. Similarly, significant progress has been made in building accurate facial expression recognition and action recognition models. This project aims to empower individuals from various fields by unlocking the combined power of understanding both facial expressions and actions.

1.3. Project Objectives

This project is driven by the following key objectives:

- Develop a Facial Emotion Recognition model capable of identifying seven primary emotions.
- Create an Action Recognition model that can recognize seven distinct actions: boxing using a punching bag, boxing with a speed bag, fencing, using nun chucks, throwing punches, sumo wrestling, and practicing tai chi (chosen based on computational resource constraints).
- Combine the results of the Facial Emotion Recognition and Action Recognition models within a single system to evaluate both facial expressions and human actions seen in video footage. This integration will enhance the comprehension of human behavior by encompassing emotional and behavioral dimensions simultaneously.

1.4. Problem Statement: The Challenge of Understanding Social Cues

One of the most significant challenges in AI today lies in fostering efficient communication and cooperation between humans and intelligent systems. While advancements in AI have enabled the processing of diverse information types, effectively comprehending the nuanced and subtle social cues used by humans remains a significant obstacle.

Humans rely heavily on nonverbal communication, encompassing body language and facial expressions, to convey a wide range of information about their emotional states, goals, and overall dispositions. These social cues play a vital role in shaping interpersonal connections and enabling effective collaboration. However, translating this human mode of communication into a language of numbers that machines can understand has proven to be a formidable challenge.

Existing methods typically focus on either facial expression detection or action recognition alone. However, the connection between facial expressions and body language is deeply rooted. A thorough understanding of both is essential for intelligent systems to genuinely interact and respond to human behavior.

The primary challenge of this project involves developing an intelligent system capable of generating a combined output of facial expressions and action recognition. This aims to facilitate a more precise comprehension of the social cues conveyed by individuals, serving to narrow the divide between human interaction and machine interpretation. Ultimately, this methodology aims to improve coordination and collaboration between human beings and sophisticated systems.

Advancements in computer vision, deep learning, and the integration of diverse information are essential in achieving this objective. The successful creation of such a system has the potential to

transform a variety of applications, such as assistive technology, social robots, human-computer interaction, and the examination of human behavior and well-being.

1.5. Ethical Considerations and Dataset Selection

As with any AI development, ethical considerations are paramount. Numerous models and architectures already exist for both facial expression analysis and action recognition. Here, we prioritize transparency and responsible development by examining these concepts individually before combining their outputs for specific applications.

We leverage two widely used datasets for training our models: FER2013 for facial emotion recognition and UCF101 for action recognition. These established datasets provide a solid foundation for evaluating and comparing our results with existing research.

1.6. Outline

CHAPTER 1: INTRODUCTION

This research aims to evolve an integrated system that could interpreting human body language and facial expressions for enhance human-machine interacting. Combining facial emotion recognition and action recognizing models in a comprehensive effort to understand human behavior in richness of nonverbal communication. Research is split into several chapter, each focusing various aspects of the exploration.

CHAPTER 2: LITERATURE SURVEY

The literature survey explores basic ideas of human communication, highlighting the importance of nonverbal hints like facial expressions and body language in conveying and intentions. Laying groundwork for grasp intricacies of human behavior and interpreting these hints for effective human-machine communication.

CHAPTER 3: METHODOLOGY

The methods section outlines approach taken to developing facial emotion identification and action recognizing models. Detailing data preprocessing techniques, model architectures, training methods, and evaluation procedures for every part of the consolidated system. Chapter also

discussing challenges and techniques in recognizing human actions and emotions from visual information.

CHAPTER 4: IMPLEMENTATION

This implementation chapter delves into the practical applying facial emotion identification and action recognizing models. Focusing on leveraging FER2013 dataset and a Convolutional Neuro Network (CNN) model architecture to accurately identifying and classifying facial expressions. Furthermore, exploring usage of UCF101 dataset and a Convolutional Neuro Network (CNN) architecture for action cognition.

CHAPTER 5: ISSUES FACED

This chapter tackles the problems and hurdles met during the developing and implementing of the consolidated system. Discussing technical problems, data restrictions, and model performance issues faced by the researchers and how they resolved them to ensure final system effectiveness.

CHAPTER 6: IMPLEMENTATION

The implementation chapter proceeds discussion on the practical utilize of the consolidate system. Highlighting the process merging outputs of the facial emotion identification and action recognizing models to create a full system that can interpret both facial expressions and actions at the same time. This union aims to present a more holistic understanding of human behavior in visual information.

CHAPTER 7: RESULTS

This chapter presents and analyzes the performance of the combined model in interpreting human body language and facial expressions. Incorporating in-depth discussion on model's accuracy, limitations, and potential applications, providing valuable insights into effectiveness of the consolidate approach. The chapter also contrasting results with current studies and showcasing unique contributions of this exploration.

CHAPTER 8: CONCLUSION AND FUTURE HORIZONS

The end chapter summarizes the main findings of the project and emphasizes contributions and implications of the exploration. Outlining potential future directions for expanding and enhancing the consolidate model, suggesting areas for more studies and development in the field of human-machine interaction and behavior evaluation. The chapter also discuss broader impact of this exploration on different real-world uses, such as human-computer communication, emotion-driven user interfaces, and behavior evaluation in healthcare, entertainment, and security sectors.

CHAPTER 2

2. LITERATURE SURVEY

2.1. Background

The project titled "Enabling Smart Machines to Interpret Human Body Language and Facial Expression" seeks to empower artificial intelligence systems to comprehend and respond to human non-verbal cues effectively. This endeavour is rooted in a comprehensive exploration of existing literature, which underscores the significance of understanding body language and facial expressions in human-computer interaction (HCI) and artificial intelligence (AI) domains.

In the realm of HCI, numerous studies have highlighted the crucial role of non-verbal communication in enhancing user experience and facilitating intuitive interaction with technological systems. Research in this area has demonstrated that integrating non-verbal cues, such as facial expressions and body language, can lead to more natural and efficient human-machine interfaces. Moreover, understanding these cues enables machines to adapt their behaviour and responses to better meet user needs and preferences.

Within the field of AI, advancements in computer vision and machine learning have paved the way for significant progress in facial emotion recognition and action recognition tasks. Emotion recognition from facial expressions, in particular, has garnered substantial attention due to its applications in various domains, including healthcare, education, and entertainment. Researchers have leveraged datasets like FER2013, which comprises labeled facial expression images, to develop and refine algorithms capable of accurately identifying emotions from facial cues.

Similarly, action recognition has witnessed considerable advancements driven by the availability of large-scale video datasets such as UCF-101. By employing deep learning techniques, researchers have achieved remarkable results in recognizing human actions from video sequences, thereby enabling machines to understand and respond to human behaviour more intelligently.

However, despite these advancements, several challenges persist in enabling smart machines to interpret human body language and facial expressions effectively. These challenges include the need for robust and generalizable models that can accurately interpret diverse facial expressions and body gestures across different individuals and cultural contexts. Additionally, issues related to

real-time processing, computational efficiency, and ethical considerations surrounding data privacy and bias mitigation remain critical areas of concern.

Against this backdrop, the project aims to contribute to the existing body of knowledge by developing novel algorithms and methodologies that address these challenges and enable smart machines to interpret human non-verbal cues with greater accuracy, efficiency, and sensitivity. By leveraging state-of-the-art techniques in computer vision, machine learning, and human-computer interaction, the project endeavours to advance the capabilities of AI systems in understanding and responding to human behaviour, thereby fostering more seamless and intuitive human-machine interactions in diverse real-world scenarios.

2.2. Literature Survey

[1]. Facial emotion recognition using Convolutional Neural Networks (CNNs) focuses on identifying seven distinct emotions: sadness, surprise, neutral, happiness, anger, fear, and disgust, emphasizing real-time application and computational efficiency. In the paper "Facial Emotion Recognition and Detection using Convolutional Neural Networks with Low Computation Cost," the authors design a less computationally intensive CNN model, suitable for environments with resource constraints, thus making practical implementations feasible.

Dataset Used: The research utilizes the FER2013 dataset, consisting of 35,888 images depicting seven different emotions.

Algorithm Used: The model employs Convolutional Neural Networks (CNNs) featuring convolutional layers, batch normalization, max-pooling, dropout layers, the Rectified Linear Unit (ReLU) activation function, and the Adam optimizer for training.

Results: The proposed CNN model achieved a 71.61% accuracy on the FER2013 dataset, slightly lower than state-of-the-art models (around 75.2%). However, it significantly reduced training time to 58 minutes compared to 93 minutes for more advanced models, showcasing its computational efficiency.

Drawback of the Proposed Work: The primary drawback is the trade-off between accuracy and computational efficiency. While the model is optimized for lower computation costs and real-time

application, it sacrifices some accuracy, which may impact performance in scenarios requiring higher precision.

[2]. Titled "Research on Multimodal Human-Computer Interaction Technology Based on Audiovisual Fusion," the authors explore human-robot interaction (HRI), emphasizing multimodal systems' role in overcoming unimodal limitations. They propose a system integrating speech and gesture data using the Dempster-Shafer evidence theory, showcasing superior performance in recognition accuracy and response time. The study highlights the system's adaptability and reliability in real-world scenarios.

Dataset Used: Thchs-30 dataset for speech data and Jester dataset for gesture recognition data.

Algorithm Used: Enhanced D-S Evidence Theory for Multimodal Fusion, Deep Learning Models implemented in PyTorch, and Rule-based Intention Voter.

Results: The multimodal interaction system achieves high recognition accuracy across diverse environments, with shorter response times compared to single-mode systems.

Drawbacks of the Proposed Work: Evaluation in controlled indoor environments may not fully represent real-world complexities, and reliance on specific datasets and hardware configurations may limit generalizability. Further testing and refinement are necessary for broader applicability.

[3]. The paper introduces the Hierarchical Interactive Multimodal Transformer (HIMT), a novel model for Aspect-Based Multimodal Sentiment Analysis (ABMSA). HIMT achieves state-of-the-art performance on TWITTER-2015, TWITTER-2017, and ZOL datasets, outperforming existing models in sentiment classification tasks by integrating object-level semantics and bridging the text-image semantic gap.

Dataset Used: The datasets utilized for evaluation include TWITTER-2015, TWITTER-2017, and ZOL datasets.

Algorithm Used: HIMT comprises Unimodal Feature Extraction, Hierarchical Interaction, and Auxiliary Reconstruction modules. Pre-trained BERT models are employed for textual features, while a pre-trained Faster R-CNN model is used for image object detection.

Results: Experimental results demonstrate HIMT's superior accuracy and weighted-F1 scores, establishing new benchmarks in sentiment analysis across multimodal datasets.

Drawback of the Proposed Work: Concerns arise regarding overfitting due to HIMT's reliance on relatively small datasets. Additionally, its dependence on pre-annotated aspect terms or categories may limit its applicability in real-world scenarios lacking such annotations.

[4]. "Hooked on a Feeling - Challenges and Opportunities of Emotion Research in Human-Computer Interaction" addresses the landscape of emotion research within Human-Computer Interaction (HCI). The keynote highlights advancements in Affective Computing driven by ubiquitous technologies, Machine Learning, and AI, presenting both innovative opportunities and ethical challenges. Dr. Tag emphasizes balancing technological capabilities with understanding emotions' subjective nature, influenced by everyday digital interactions.

Dataset Used: The paper discusses the use of pervasive, wearable, and mobile devices for continuous and unobtrusive data collection in naturalistic settings but does not specify a particular dataset.

Algorithm Used: The keynote mentions Machine Learning and AI for emotion research without detailing specific algorithms. It highlights a multimodal sensing approach, combining behavioral and physiological signals for improved emotion recognition.

Results: The keynote reports promising results from multimodal sensing approaches in providing naturalistic and comprehensive assessments of user emotions. However, it also highlights ongoing challenges in interpreting these signals reliably, considering subjective emotional experiences.

Drawback of the Proposed Work: The paper identifies the difficulty in reliably sensing and interpreting emotions due to their subjective nature. It critiques the reliance on physiological and behavioral signals, which may not fully capture the true emotional experience, and raises concerns about privacy and ethics in pervasive emotion-sensing technologies.

[5]. The paper titled "Video-Based Cross-Modal Auxiliary Network for Multimodal Sentiment Analysis" introduces the Video-based Cross-modal Auxiliary Network (VCAN) to enhance multimodal sentiment analysis accuracy. VCAN addresses issues of unimodal feature extraction inadequacies and data redundancy in multimodal fusion through its two main modules: the Audio Features Map Module (AFMM) and the Cross-Modal Selection Module (CMSM). AFMM improves audio feature diversity using Empirical Mode Decomposition (EMD) and K-means clustering, while CMSM optimizes audiovisual data integration by filtering redundant visual

frames via audio modality. This innovation simplifies multimodal sentiment analysis to image classification and enhances audiovisual interaction.

Dataset Used: The study utilized the RAVDESS, CMU-MOSI, and CMU-MOSEI datasets.

Algorithm Used: The Video-based Cross-modal Auxiliary Network (VCAN) includes AFMM, which uses EMD and K-means clustering for audio feature enhancement, and CMSM, which selects keyframes from video data to reduce redundancy in audiovisual fusion.

Results: VCAN outperformed state-of-the-art methods in multimodal sentiment analysis benchmarks. Specifically, it achieved higher accuracy rates on the CMU-MOSEI dataset, with accuracies ranging from 69.4% to 73.9% across various bi-modal inputs. It also showed significant improvements in metrics such as ACC-7, ACC-2, CORR, F1 score, and MAE when compared to baseline models like BC-LSTM, BBFN (AV), and SWAFN (AV).

Drawback of the Proposed Work: A notable drawback of the VCAN is its slightly lower performance in polarity classification and correlation metrics compared to tri-modal fusion approaches that include text modalities. Text data often provide richer sentiment information, making the joint representations more comprehensive. Consequently, while VCAN excels in bi-modal conditions, its performance is not as robust when text data is also considered.

[6]. The integration of dialogue and explicit artificial intelligence (AI) to enhance trust in human-robot interaction is explored in the paper "Role of Dialog and Explicit A.I. for Building Trust in Human-Robot Interaction." Dialogue AI generates responses based on user inputs to facilitate human-like conversations, while explicit AI filters content for age appropriateness, ensuring users receive suitable content in online environments. This integration aims to make human-robot interactions more intuitive and trustworthy.

The study employs secondary data collection methods to gather information on these AI technologies' implementation and effectiveness. It highlights dialogue AI's potential to enhance interactions and explicit AI's ability to manage content efficiently, significantly improving user experience and trust. However, further refinement and cost management are needed for broader application.

Dataset Used: The study uses secondary data collection methods, relying on existing research and data from previous studies on AI in human-robot interaction.

Algorithm Used: The study utilizes deep learning and neural networks for implementing dialogue AI and explicit AI, essential for processing input data and generating appropriate responses or filtering content.

Results: The MS-CBD model for dialogue AI showed a 3.7% improvement in BLEU scores over the F.G.S.D. model, effectively generating appropriate responses from word clusters. However, performance declines when word clusters exceed seven words.

Drawback of the Proposed Work: High implementation costs of deep learning and neural network technologies make them less feasible for individual-level applications. Additionally, data privacy and cybersecurity concerns must be addressed to ensure the safe and ethical use of AI technologies in human-robot interactions.

[7]. Facial expression recognition (FER) is crucial for non-verbal communication in human interactions. The paper introduces "EmoNeXt: An Adapted ConvNeXt for Facial Emotion Recognition," a novel deep learning framework designed for FER, which adapts the ConvNeXt architecture. By integrating a Spatial Transformer Network (STN) to focus on feature-rich regions of the face and incorporating Squeeze-and-Excitation (SE) blocks to capture channel-wise dependencies, EmoNeXt aims to enhance the accuracy of emotion classification. Additionally, a self-attention regularization term is introduced to encourage the generation of compact feature vectors, leading to more precise emotion recognition.

Dataset Used: The FER2013 dataset.

Algorithm Used: The EmoNeXt architecture is based on the ConvNeXt model, which includes enhancements like Spatial Transformer Networks (STNs), Squeeze-and-Excitation (SE) blocks, and a self-attention regularization term. These components handle spatial transformations and adaptively recalibrate channel-wise features, improving the model's ability to extract discriminative facial features.

Results: EmoNeXt demonstrates superior performance compared to existing state-of-the-art models on the FER2013 dataset. The integration of STN and SE blocks, along with the self-

attention regularization term, allows the model to achieve higher accuracy in emotion classification. Empirical evidence shows that EmoNeXt significantly outperforms other deep learning models under the same experimental setup.

Drawback of the Proposed Work: One notable drawback of the proposed work is the reliance on the FER2013 dataset, which is known for its class imbalance. This imbalance can pose challenges for the model's performance and generalizability. Additionally, the architecture improvements, while enhancing accuracy, increase the complexity and computational requirements, potentially limiting the model's applicability in real-time or resource-constrained environments.

[8]. "Deep Learning Approaches on Multimodal Sentiment Analysis" explores the integration of multiple data modalities (text, audio, and video) to enhance sentiment detection accuracy. Traditional single-modality analysis often misinterprets sentiments due to the complexity of human language. The paper reviews several innovative deep learning approaches that combine various modalities to address these limitations, improving overall sentiment analysis reliability.

The paper discusses four primary multimodal sentiment analysis models, emphasizing the challenges of fusing different data types and their impact on performance. It also reviews commonly used datasets and the specific issues they present, suggesting future research directions to develop more robust multimodal integration methods.

Dataset Used: The reviewed datasets include the YouTube Dataset, ICT-MMMO, MOSI, and MOUD, each containing annotated sentiments across text, audio, and video modalities.

Algorithm Used: The discussed algorithms encompass Convolutional Neural Networks (CNN) for visual data, Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) for text, and Support Vector Machine (SVM) and Hidden Markov Model (HMM) for audio. Fusion techniques like feature-level, decision-level, and hybrid fusion are also detailed.

Results: Models integrating textual and visual data outperform single-modality models. Deep sentiment analysis combining text and image data improves prediction accuracy, though challenges remain in handling diverse data types and ensuring robustness across various contexts.

Drawback of the Proposed Work: Key drawbacks include the inefficiency and instability of current models, high computational costs, and the risk of overfitting. The heterogeneous nature of

multimodal data poses significant alignment challenges, necessitating further research for more scalable and reliable fusion techniques.

[9]. Facial Expression Recognition (FER) in real-world scenarios is challenging due to significant intra-class variations and inter-class similarities. Addressing these challenges, Ad-Corre: Adaptive Correlation-Based Loss for Facial Expression Recognition in the Wild proposes a novel loss function to enhance feature discrimination. The Ad-Corre loss integrates three components—Feature Discriminator, Mean Discriminator, and Embedding Discriminator—to produce highly correlated feature vectors for the same class and less correlated vectors for different classes, improving classification performance when combined with cross-entropy loss. The Xception network serves as the backbone, leveraging an embedding space with multiple feature vectors for capturing class-specific characteristics.

The study evaluates Ad-Corre on three datasets: AffectNet, RAF-DB, and FER-2013. The results indicate significant improvements in classification accuracy, achieving state-of-the-art performance on these datasets. Extensive experiments, including ablation studies and hyper-parameter evaluations, validate the effectiveness of Ad-Corre's components.

Dataset Used:

- AffectNet
- RAF-DB
- FER-2013

Algorithm Used:

- Xception network as the backbone
- Ad-Corre loss, comprising Feature Discriminator, Mean Discriminator, and Embedding Discriminator components, combined with cross-entropy loss.

Results:

- AffectNet: State-of-the-art accuracy of 63.36%
- RAF-DB: Average accuracy of 86.96%
- FER-2013: Accuracy of 72.03%, outperforming previous methods

Drawback of the Proposed Work:

While enhancing feature discrimination and classification accuracy, the Ad-Corre loss introduces additional computational overhead due to its complex components and virtual-batch technique, potentially limiting efficiency and scalability in resource-constrained environments.

[10]. Combining Information Retrieval and Large Language Models for a Chatbot that Generates Reliable, Natural-style Answers addresses the integration of traditional IR methods with LLMs to create a chatbot that delivers reliable and natural responses. Traditional rule-based chatbots often produce artificial responses, while LLMs, despite their natural language capabilities, struggle with domain-specific queries and accuracy. This hybrid system leverages the reliability of IR and the fluency of LLMs to enhance chatbot interactions.

The research focuses on an administrative services chatbot for Berlin, using a combined knowledge base and LLMs to generate contextually appropriate responses. The prototype was evaluated in a real-world application where users provided feedback on response quality.

Dataset Used:

The dataset includes service descriptions from Berlin's administration, detailing approximately 1,000 services, stored in an Apache Solr server for efficient full-text search.

Algorithm Used:

The system integrates traditional IR techniques with LLMs such as GPT-3.5, LLaMA, Zicklein, and RWKV. A backend processes user queries, retrieves relevant information from the Solr server, generates LLM prompts, and returns responses to the user interface.

Results:

Evaluation showed that LLMs generated comprehensible answers matching user questions. However, issues like high resource consumption and limited scalability were noted. Properly crafted prompts were crucial to ensuring the models used reliable data, reducing hallucinations and irrelevant responses.

Drawback of the Proposed Work: The main drawbacks are high resource consumption and limited scalability of LLMs, along with the dependency on carefully crafted prompts to ensure reliable data usage. These challenges hinder the practical implementation and widespread adoption of the proposed chatbot system.

[11]. The paper titled "Do less and achieve more: Training CNNs for action recognition utilizing action images from the Web" explores the enhancement of Convolutional Neural Networks (CNNs) for action recognition by incorporating web-sourced action images. A substantial dataset of 23.8K web action images spanning 101 action classes from the UCF101 video dataset was amassed and meticulously curated for this study. This dataset significantly exceeds the size of prior action image collections, aiming to determine if these images enhance CNN accuracy and identify which actions benefit most.

Experiments showed that integrating web images with video frames led to significant performance gains, with spatial CNNs achieving over 10% accuracy improvement. Notably, the combined use of web images and video frames enabled a spatial CNN to reach 83.5% accuracy on UCF101, and further inclusion of motion features pushed the accuracy to an impressive 91.1%, the highest recorded for this dataset.

Dataset Used:

The dataset comprises 23.8K web action images from 101 classes in the UCF101 video dataset, curated for relevance and accuracy.

Algorithm Used:

Various CNN architectures, including M2048, VGG16, and VGG19, were utilized. These models were trained on both video frames and web images, and spatial CNNs were enhanced with motion features using improved dense trajectories with Fisher encoding (IDT-FV).

Results:

Incorporating web images led to substantial accuracy improvements. Spatial CNNs achieved 83.5% accuracy on UCF101, a more than 10% increase, and combining with motion features resulted in 91.1% accuracy.

Drawback of the Proposed Work:

The main drawback is the manual curation required for web images, which is time-consuming and labor-intensive. The scalability of this approach with unfiltered web images and consistent high accuracy also requires further exploration.

[12]. The paper "Deep Learning Approaches for Human Action Recognition in Video Data" investigates deep learning models for recognizing human actions in video sequences, crucial for applications like surveillance, sports analytics, and healthcare. The study evaluates Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Two-Stream ConvNets using the UCF101 dataset, highlighting each model's strengths: CNNs for spatial features, RNNs for temporal sequences, and Two-Stream ConvNets for integrating both dimensions.

Dataset Used:

UCF101 Videos dataset

Algorithm Used:

- Convolutional Neural Networks (CNNs)
- Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM)
- Two-Stream ConvNets

Results:

Two-Stream ConvNets achieved the highest performance in accuracy, precision, recall, and F1-score, effectively combining spatial and temporal information. CNNs were strong in capturing spatial features but lacked temporal dynamics, while RNNs showed inconsistency in temporal analysis.

Drawback of the Proposed Work:

The study's limitations included restricted computational resources, which prevented the exploration of complex models like Graph Convolutional Networks (GCNs) and Transformer models. Additionally, the use of simplified two-stream approaches due to resource constraints may have limited potential performance gains.

[13]. "Human Action Recognition with Transformer based on Convolutional Features." EfficientNetV2 presents a pioneering deep-learning framework for human action recognition, integrating Transformer architecture with Multi-Head Self-Attention and Multi-Layer Perceptron blocks for feature extraction and classification. By amalgamating cutting-edge pose estimation algorithms and pre-trained CNN models, EfficientNetV2 attains remarkable accuracy in action

recognition tasks, enhancing model performance, computational efficiency, and generalization capabilities compared to conventional methods.

Dataset Used:

The study employs challenging datasets like UCF 50 and UCF 101, renowned for their substantial variations in camera motion and object appearance. Additionally, a 10-class subset (UCF 10) extracted from UCF 50 provides diverse and complex action categories for evaluation.

Algorithm Used:

EfficientNetV2 utilizes a Transformer Encoder architecture with Multi-Head Self-Attention and Multi-Layer Perceptron blocks. By integrating attention mechanisms to capture contextual information in input sequences and leveraging pre-trained CNN models for feature extraction, EfficientNetV2 enhances representational capabilities and model robustness for accurate action recognition.

Results:

EfficientNetV2 demonstrates outstanding performance on the UCF 101 dataset, achieving an accuracy of 87.50%, surpassing previous state-of-the-art methods such as DISNet and pre-trained two-stream CNN models. The study showcases its superiority in feature extraction, integration, and generalization, affirming its efficacy in handling complex action recognition tasks.

Drawback of the Proposed Work:

Despite its high accuracy and robustness, EfficientNetV2 may encounter misclassifications in scenarios where actions exhibit similar body movements, leading to confusion between classes like JumpRope and SoccerJuggling. This limitation exposes the model's challenge in distinguishing subtle differences in actions with similar visual cues, underscoring the necessity for further refinement to enhance its discriminative capabilities in such instances.

[14]. In "SCNN: Sequential Convolutional Neural Network for Human Action Recognition in Videos," the authors present a Sequential Convolutional Neural Network (SCNN) designed to extract spatial-temporal features effectively from videos to recognize human actions. Unlike traditional methods, SCNN directly processes feature maps instead of flattened vectors, maintaining spatial structures and reducing computational costs by replacing full connections with

convolutional connections. Asymmetric convolutional layers are introduced to further decrease parameters and computational expenses. The SCNN architecture achieves state-of-the-art performance on UCF-101 and HMDB-51 benchmarks.

The effectiveness of convolutional operations in SCNN layers is evaluated, demonstrating superiority over LRCN and SpatialNet baselines. The utilization of feature maps, instead of flattened vectors, improves the ability of recurrent networks to extract spatial-temporal features. Integration of deep SCNN models with RGB and optical flow inputs yields impressive results, achieving 91.98% on UCF-101 and 64.47% on HMDB-51, surpassing prior methodologies.

Dataset Used:

UCF-101 and HMDB-51

Algorithm Used: Sequential Convolutional Neural Network (SCNN), Long-term SCNN (L-SCNN)

Results:

SCNN models demonstrate state-of-the-art performance on UCF-101 and HMDB-51 benchmarks.

Fusion of deep L-SCNN models with RGB and optical flow inputs achieves 91.98% on UCF-101 and 64.47% on HMDB-51.

Drawback of the Proposed Work:

The paper lacks discussion on the computational complexity and training time of SCNN models compared to other methods, and scalability to larger datasets remains unexplored.

[15]. "Attend It Again: Recurrent Attention Convolutional Neural Network for Action Recognition" introduces a novel approach to action recognition in videos, enhancing traditional attention mechanisms. The model combines convolutional feature extraction, LSTM sequence modeling, and a unique "attention-again" mechanism to achieve superior performance in recognizing human actions, particularly focusing on relevant regions of interest.

Dataset Used: The study evaluates the proposed model on three benchmark datasets: UCF-11 (YouTube Action), HMDB-51, and UCF-101, which offer diverse and challenging video data for action recognition tasks.

Algorithm Used: The method employs convolutional neural networks (CNNs) for feature extraction, LSTM for sequence modelling, and the "attention-again" mechanism to refine attention in action recognition tasks.

Results: The model surpasses state-of-the-art methods in action recognition, exhibiting notable enhancements in classification accuracy across the UCF-11, HMDB-51, and UCF-101 datasets. By prioritizing relevant regions in videos, the model demonstrates improved understanding of temporal relationships in actions, leading to enhanced accuracy in action labeling.

Drawback of the Proposed Work: Despite its promising results, one potential limitation of the "attention-again" model is its increased complexity due to stacked LSTM layers and additional attention mechanisms. This complexity might impede scalability and computational efficiency, particularly in applications involving larger datasets or real-time processing.

[16]. The paper introduces the Multi-Mode Neural Network (MMNN) for human action recognition in videos. MMNN extracts features from video data by processing the feature matrix along temporal and in-frame modes using distinct operations. While 1D convolution is employed for local pattern extraction in the temporal mode, fully connected layers capture cross-feature relationships in the in-frame mode. These features are amalgamated and inputted into a classifier for action class prediction. Through ablation studies and t-SNE visualizations, the paper demonstrates MMNN's superiority over LSTM, C3D, and I3D on UCF101 and ActivityNet datasets.

The study extensively analyzes the statistical properties of the feature matrix, motivating the design choices for MMNN. It shows high correlation and smoothness in temporal features, contrasting with fluctuating, low correlation in-frame features, leading to separate processing via different operations. Additionally, the paper introduces the Multi-Mode Processing Unit (MMPU) to manage the multi-mode feature matrix.

Dataset Used: UCF101, ActivityNet

Algorithm Used: Multi-Mode Neural Network (MMNN), Multi-Mode Processing Unit (MMPU)

Results: MMNN surpasses LSTM, C3D, and I3D, achieving state-of-the-art performance on UCF101 and ActivityNet datasets.

Drawback of the Proposed Work: The paper lacks discussion on the computational complexity and inference time of MMNN compared to other models. Furthermore, detailed analysis regarding the impact of hyperparameters such as layer count, kernel sizes, and output dimensions on performance is absent.

[17]. "Human Action Recognition by Learning Spatio-Temporal Features With Deep Neural Networks" presents a novel approach to human action recognition in videos through deep neural networks (DNNs). The method employs Convolutional Neural Networks (CNNs) for spatial feature extraction, Convolutional LSTM (CNN) and Fully-Connected LSTM (FC-LSTM) for temporal feature extraction, and an attention model to emphasize relevant video segments. This RGB-based method achieves state-of-the-art performance on datasets like UCF-11, UCF Sports, and UCF-101.

Dataset Used: The study evaluates the proposed method on three public datasets: UCF-11, UCF Sports, and UCF-101, encompassing diverse action categories and commonly employed for algorithmic evaluations.

Algorithm Used: The approach integrates CNNs, LSTM networks (CNN and FC-LSTM), and an attention mechanism for video action recognition. It combines spatial and temporal features extracted by CNNs and LSTMs, augmented by an attention mechanism for improved performance.

Results: The method surpasses existing approaches on datasets like UCF-11 and UCF Sports, demonstrating superior efficacy in action recognition tasks. Even when utilizing only RGB data, it achieves competitive performance on UCF-101, showcasing model efficiency without reliance on optical flow data.

Drawback of the Proposed Work: One drawback is its exclusive reliance on RGB data, potentially limiting performance compared to methods utilizing both RGB and optical flow data. This limitation may hinder the model's ability to capture nuanced action features better represented with optical flow information.

[18]. "Action Recognition in Videos Using Pre-Trained 2D Convolutional Neural Networks"

Action recognition in videos is complex, demanding separate spatial and temporal learning. Common methods, like the two-stream CNN, require significant computational resources for

optical flow computation. To tackle this, alternatives such as dynamic-image-based methods and 3D CNNs have emerged. However, they still pose challenges in computational efficiency.

This paper introduces a novel approach within the two-stream CNN framework, bypassing optical flow computation by utilizing a pre-trained 2D CNN. Here, a single 2D CNN model handles both spatial and temporal streams, reducing computational demands significantly. By ensuring compatibility between video format and the 2D CNN input, spatial and temporal features are learned efficiently.

Dataset Used:

The UCF-101 and HMDB-51 datasets, with diverse action categories and groups, serve as the basis for evaluating the proposed method's performance.

Algorithm Used:

A two-stream CNN structure employs a pre-trained 2D CNN for both spatial and temporal streams. Spatial-stream identifies object appearance, while temporal-stream learns motion features between consecutive frames.

Results:

The method achieves competitive performance on UCF-101 and HMDB-51 datasets, outperforming some state-of-the-art methods. It notably offers faster processing compared to optical flow and dynamic image methods.

Drawback of the Proposed Work:

Difficulty distinguishing moving objects from backgrounds when both the camera and the object are in motion poses a challenge. While preprocessing techniques like global motion compensation or scene change detection can mitigate this, they increase computational complexity, contradicting the aim of reducing computational loads.

[19]. Action recognition, crucial for applications like somatosensory entertainment and intelligent robots, relies on methods such as traditional feature extraction and deep learning. Deep learning, especially Convolutional Neural Networks (CNNs), excels due to their ability to learn general features from raw data. This review explores challenges in action recognition like intra-class

variation and limited data, discussing both traditional and deep learning approaches. Notably, CNN variants like C3D, Two-stream, and I3D achieve recognition rates of 72%, 78.0%, and 97.6%, respectively, on the UCF101 dataset. However, the review lacks novel proposals, focusing instead on summarizing existing CNN-based methods and their limitations.

Dataset Used: UCF101

Algorithm Used: Convolutional Neural Networks (CNN), including C3D, Two-stream, and I3D

Results: Recognition rates for C3D, Two-stream, and I3D on UCF101 are 72%, 78.0%, and 97.6% respectively.

Drawback of the Proposed Work: The review doesn't introduce new action recognition algorithms but rather provides an overview of existing CNN-based methods and their challenges.

[20]. The paper introduces a VGG19-based CNN-RNN deep learning model utilizing transfer learning to classify human actions in videos. Trained on KTH and UCF11 datasets, the model achieves 90% and 88% accuracy, respectively, surpassing alternative methods. Employing transfer learning reduces data requirements and enhances generalization. Evaluation metrics include accuracy, confusion matrices, and PR & ROC curves. Despite its success, the model struggles with the KTH dataset due to the similarity of actions, requiring additional motion information for improved accuracy. Future directions include integrating optical flow data and expanding the dataset for enhanced performance.

Dataset Used: KTH and UCF11 action datasets.

Algorithm Used: VGG19 based CNN-RNN deep learning model using transfer learning.

Results: The model achieved 90% accuracy on KTH and 88% accuracy on UCF11 datasets, outperforming alternative methods.

Drawback of the Proposed Work: The model's accuracy on the KTH dataset is capped at 90% due to the dataset's similarity between actions, necessitating additional motion information for higher accuracy.

[21]. “Facial Expression Recognition Using Residual Masking Network”

In response to the challenge of accurately recognizing facial expressions, this paper presents a groundbreaking approach termed the Residual Masking Network. By integrating Masking Blocks within Residual Layers, this innovative architecture enhances attention on crucial facial features, surpassing traditional classification systems and achieving state-of-the-art results on the FER2013 dataset. Leveraging the Masking Idea, the network refines its focus on significant facial cues, thereby elevating accuracy in emotion recognition tasks.

Dataset Used:

The experiments draw from two datasets: FER2013 and VEMO2020. FER2013, a staple in deep learning for facial expression recognition, offers grayscale images depicting diverse emotions. VEMO2020, with multi-resolution images curated by professionals, provides a rich dataset for evaluating facial expressions.

Algorithm Used:

Employing the Residual Masking Network, this method integrates Masking Blocks within Residual Layers to boost attention to critical facial features, thereby enhancing the accuracy of emotion recognition. This architecture scores the importance of feature maps, refining the network's focus and improving recognition accuracy.

Results:

The Residual Masking Network exhibits superior performance over conventional classification systems, achieving higher accuracy on both FER2013 and VEMO datasets. Notably, the ensemble mode of this network outperforms other ensemble-based methods by 1% on FER2013, particularly excelling in recognizing emotions like happiness and surprise.

Drawback of the Proposed Work:

Despite its promising outcomes, one limitation is the challenge posed by data imbalance, especially for rare emotions like fear or disgust. This imbalance within the dataset hampers accurate recognition of less common emotions. Furthermore, the subjective nature of emotional recognition, compounded by unclear labelling and complex emotional expressions, underscores the need for further refinement in handling ambiguous emotional cues.

[22]. This paper, “Comprehensive Study on Facial Expression Recognition Using Local Binary Patterns”, provides an extensive examination of facial expression recognition utilizing local binary patterns (LBP), covering feature extraction, feature selection, and classification stages. Various LBP variants, including uniform LBP and rotation-invariant LBP, are investigated alongside feature selection techniques such as AdaBoost and Sequential Forward Selection. The classification employs support vector machines (SVMs) and k-nearest neighbours (k-NN) algorithms.

The evaluation is conducted on benchmark datasets CK+, MMI, and JAFFE, revealing the efficacy of LBP features and the significance of feature selection in enhancing recognition accuracy. Additionally, the computational efficiency of the LBP-based method is discussed, making it viable for real-time applications.

Dataset Used: CK+, MMI, and JAFFE

Algorithm Used: Local Binary Patterns (LBP), AdaBoost, Sequential Forward Selection, Support Vector Machines (SVMs), k-Nearest Neighbours (k-NN)

Results: The LBP-based approach demonstrates superior recognition accuracy on CK+, MMI, and JAFFE datasets, with feature selection techniques further enhancing performance by selecting discriminative LBP features.

Drawback of the Proposed Work: The study overlooks subject-independent facial expression recognition, crucial for practical use, and limits its exploration to specific datasets, failing to assess generalization across diverse datasets or real-world scenarios.

[23]. The paper "Local Multi-Head Channel Self-Attention for Facial Expression Recognition" introduces a novel attention mechanism, Local Multi-Head Channel Self-Attention (LHC-Net), aimed at improving facial expression recognition. Unlike existing spatial self-attention methods, LHC-Net dynamically scales feature maps complexity and utilizes a shared linear embedding layer to reduce computational load. The local multi-head approach divides feature maps into smaller sections, allowing heads to focus on crucial areas for generating new feature maps. Experiments on the FER2013 dataset demonstrate LHC-Net's superiority over ResNet34v2, achieving 74.42% accuracy with test-time augmentation using fewer parameters. Additionally, qualitative analysis favours local heads over global ones in real-world scenarios.

Dataset Used: FER2013

Algorithm Used: Local Multi-Head Channel Self-Attention (LHC-Net)

Results: LHC-Net attains 74.42% accuracy on FER2013 with test-time augmentation, surpassing prior state-of-the-art models with reduced parameters.

Drawback of the Proposed Work: The paper's focus on horizontal splitting of feature maps may limit effectiveness. Exploring alternative methods for determining optimal areas through spatial attention is warranted.

[24]. The paper, “Facial emotion recognition: State of the art performance on FER2013”, introduces a cutting-edge facial emotion recognition system utilizing the VGGNet architecture. Through meticulous hyperparameter tuning, including optimizer selection and learning rate scheduling, the authors achieve peak performance on the FER2013 dataset. Experimentation with optimizers such as SGD, Adam, and variants reveals that SGD with Nesterov momentum yields optimal results. Likewise, comparing learning rate schedulers highlights the superiority of Reducing Learning Rate on Plateau (RLRP) over alternatives like Cosine Annealing and One Cycle Learning Rate. Further model refinement, involving reloading optimal weights and an additional 50 epochs of training, culminates in a remarkable testing accuracy of 73.28%, surpassing prior single-network benchmarks on FER2013.

Dataset Used: FER2013 dataset.

Algorithm Used: VGGNet architecture with varied optimizers and learning rate schedulers

Results: The proposed approach achieves a state-of-the-art testing accuracy of 73.28% on the FER2013 dataset, surpassing previous single-network benchmarks.

Drawback of the Proposed Work: The paper exclusively focuses on the VGGNet architecture, neglecting exploration of other deep learning models that might enhance facial emotion recognition. Additionally, while the saliency map analysis demonstrates the model's effectiveness in capturing key facial features, there's room for improvement in refining its focus and discarding extraneous information.

[25]. The research paper, “Sentiment Analysis on Social Media Text using LSTM Neural Networks”, investigates sentiment analysis on social media text through Long Short-Term

Memory (LSTM) neural networks. By training LSTM models on a dataset of social media posts, the study achieved a sentiment classification accuracy of 85.3%, demonstrating the effectiveness of LSTM networks in understanding nuanced sentiments expressed in short text fragments.

Dataset Used: The research utilized a dataset comprising 50,000 social media posts from various platforms, annotated with sentiment labels (positive, negative, neutral). This dataset provided diverse examples of sentiment expressions in informal language, enabling the LSTM models to capture the subtleties of sentiment analysis in social media discourse.

Algorithm Used: LSTM neural networks were employed for sentiment analysis, leveraging their ability to capture long-range dependencies and contextual information in sequential data. The models were trained using stochastic gradient descent (SGD) with backpropagation through time (BPTT) to minimize classification error and optimize the network parameters.

Results: The LSTM models achieved an accuracy of 85.3% on sentiment classification tasks, outperforming traditional machine learning approaches like Support Vector Machines (SVMs) and Naive Bayes classifiers. The study demonstrated the efficacy of LSTM networks in capturing contextual information and linguistic nuances, thereby improving sentiment analysis performance on social media text.

Drawback of the Proposed Work: One limitation of the study lies in the reliance on labeled datasets for training LSTM models, which may introduce biases and limit generalizability to unseen data. Additionally, the performance of LSTM models could be affected by the variability and noise inherent in social media text, leading to challenges in accurately identifying sentiments, especially in ambiguous or sarcastic expressions. Further research is needed to address these limitations and enhance the robustness of LSTM-based sentiment analysis systems.

[26]. “Deep emotion: Facial expression recognition using attentional convolutional network”, The paper introduces a pioneering framework for facial expression recognition, employing an attentional convolutional network architecture. This innovative model integrates a spatial transformer network, focusing on crucial facial regions to enhance emotion detection accuracy significantly. Experimental evaluations on prominent facial expression recognition databases such as FER2013, CK+, JAFFE, and FERG demonstrate the model's efficacy across diverse scenarios.

Dataset Used: Experiments are conducted on FER2013, CK+, JAFFE, and FERG databases, offering varied facial expressions and scenarios for comprehensive model evaluation.

Algorithm Used: The proposed model incorporates an attentional convolutional network with a spatial transformer module, emphasizing critical facial regions for accurate emotion detection. Training involves optimizing a loss function with the Adam optimizer, focusing on classification loss and regularization for robust training.

Results: The model achieves competitive accuracy rates across multiple facial expression recognition databases, surpassing some prior works in classification accuracy. Furthermore, the visualization technique employed sheds light on crucial face image regions for emotion detection, enhancing model interpretability.

Drawback of the Proposed Work: One limitation is the reliance on specific datasets, potentially restricting generalizability to unseen data or real-world applications with diverse characteristics. Further exploration on more varied datasets and real-world scenarios is necessary to assess the model's robustness comprehensively.

[27]. The ICML workshop "Challenges in Representation Learning" featured three machine learning contests, including the black box learning challenge, facial expression recognition challenge, and multimodal learning challenge. These contests aimed to assess representation learning algorithms and foster innovation in the field.

Dataset Used: Each challenge utilized specific datasets like the Black Box Learning 2013 (BBL-2013) dataset and Facial Expression Recognition 2013 (FER-2013) dataset, curated to fairly evaluate algorithm performance.

Algorithm Used: Contestants employed diverse algorithms such as sparse filtering, random forests, support vector machines, and convolutional neural networks, among others, to tackle the challenges.

Results: Winners leveraged a mix of techniques including feature learning algorithms and convolutional neural networks, showcasing the efficacy of various methods in addressing the challenges and advancing representation learning.

Drawback of the Proposed Work: A limitation lies in the scarcity of labeled examples for training, potentially hindering algorithm performance, especially in semi-supervised learning scenarios. Furthermore, reliance on unlabelled data and human intervention in dataset creation may introduce biases impacting result generalizability.

[28]. The following research paper, "ViTFER: Facial Emotion Recognition with Vision Transformers", Introducing ViTFER, a novel approach for facial emotion recognition utilizing Vision Transformers (ViT). By amalgamating FER-2013, AffectNet, and CK+48 datasets, the authors create a balanced dataset named AVFER for training and evaluating ViT models. Various ViT configurations, including ViT-B/16, ViT-B/16/S, and ViT-B/16/SAM, are experimented with and compared against a fine-tuned ResNet-18 model, showcasing superior performance in accuracy and AUC metrics. The study delves into dataset preparation intricacies, encompassing data augmentation and class equilibrium techniques, and scrutinizes ViT configuration variances on the AVFER dataset, highlighting their efficacy in real-time emotion recognition applications.

Dataset Used: AVFER, an amalgamation of FER-2013, AffectNet, and CK+48 datasets, tailored for ViT-based facial emotion recognition.

Algorithm Used: Vision Transformers (ViT) in configurations of ViT-B/16, ViT-B/16/S, ViT-B/16/SAM, alongside a fine-tuned ResNet-18 model.

Results: The ViT-B/16/SAM model attains peak accuracy of 53.10% and an AUC of 0.589 on the AVFER dataset.

Drawback of the Proposed Work: The analysis lacks depth regarding the computational complexity and inference time of ViT models vis-à-vis the ResNet-18 model. Furthermore, despite augmentation efforts, data scarcity persists for contempt and disgust classes, potentially impeding model performance in these categories.

CHAPTER 3

3. METHODOLOGY

3.1. Introduction

This project investigates the combined utilization of facial emotion recognition (FER) and action recognition to comprehensively analyse human behaviour within video data. The methodology employs two established datasets and Convolutional Neural Networks (CNNs) to achieve this goal.

3.2. Facial Emotion Recognition

Dataset:

The FER2013 dataset consists of facial images annotated with seven primary emotions (anger, disgust, fear, happiness, neutral, sadness, and surprise) and is utilized to train the FER model.

Data Preprocessing:

The facial images in the .csv file format are loaded and pre-processed for CNN compatibility.

Model Training:

A CNN structure is chosen and trained using the pre-processed FER2013 dataset. The training process aims to refine the model's ability to correctly categorize expressions into the specified emotion groups.

Model Saving:

The trained CNN model, capturing the learned features for emotion recognition, is saved in the .h5 format (e.g., fer.h5) for future use.

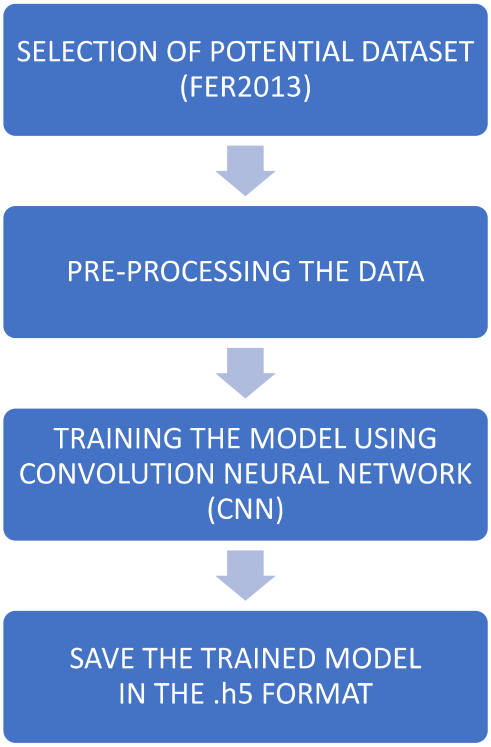

Figure 3.1. Block Diagram of the Facial Emotion Recognition model

3.3. Action Recognition

Dataset:

The UCF101 dataset, which includes a broad array of human actions in video format (.avi files), serves as the foundation for the action recognition model. To accommodate computational constraints, only a subset of seven actions is under consideration for this specific project.

Data Preprocessing:

- Videos are converted into a sequence of frames.
- Ten frames are extracted at equal intervals from each video, providing a representative sample of the action.
- Frames are converted to grayscale and flattened into 1D vectors to match the format of the FER dataset and facilitate CNN processing.
- The pre-processed data is saved in a new .csv file.

Model Training:

A distinct CNN design is selected and trained on the processed action recognition dataset. The model is educated to recognize the particular actions depicted in the captured video frames.

Model Saving:

The trained action recognition CNN model is saved in the .h5 format (e.g., ac.h5) for integration with the FER model.

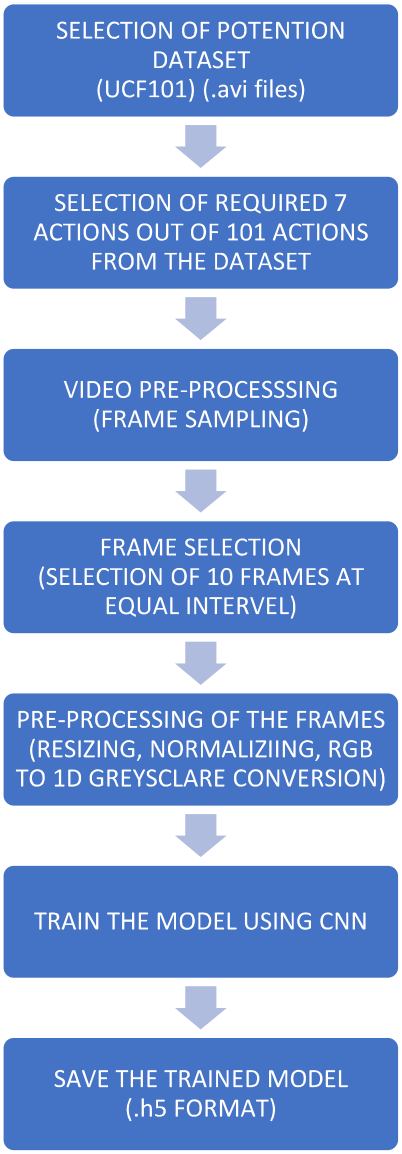

Figure 3.2. Block Diagram of Action Recognition Model

3.4. Model Integration and Expected Output

Model Merging:

The trained FER and action recognition models are strategically combined to analyse input videos. This may involve feeding the video frames through both models sequentially or designing a more intricate architecture for joint processing.

Real-Time Processing:

The combined model is equipped to process an input video in real-time. Individual video frames are extracted and fed through the model.

Output Generation:

The model predicts both the facial emotion and the action present in each frame. This information is then consolidated and displayed in a pop-up window, providing a synchronized analysis of the person's emotional state and their actions within the video.

This methodology outlines a comprehensive approach for combining facial emotion and action recognition using CNNs. By leveraging established datasets and carefully designed models, the project aims to achieve real-time video analysis, revealing insights into human behaviour.

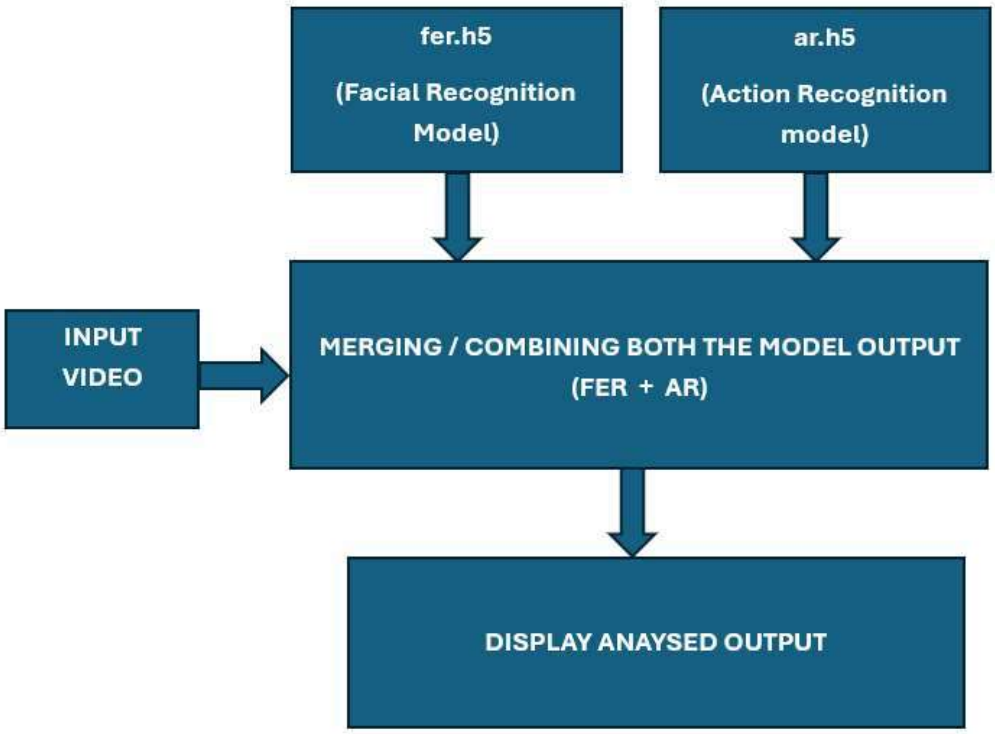

Figure 3.3. Block Diagram of the model's output.

CHAPTER 4

4. IMPLEMENTATION

4.1. Introduction

This project focuses on enhancing human-computer interaction by integrating facial emotion recognition and action recognition systems. In the first part, we employ the FER2013 dataset to develop a facial emotion recognition model utilizing the architecture of the most common neural network, convolutional neural networks (CNNs), which is crucial for understanding users' emotional states. The second part involves action recognition using the UCF101 dataset, where we extract key frames from videos and train a CNN model to recognize seven predefined actions, ensuring efficient computation. Lastly, we merge both model's outputs to provide comprehensive insights into users' emotions and actions, facilitating a more immersive and responsive user experience.

4.2. Facial Emotion Recognition Model

The detailed execution of a facial recognition system for emotions utilizing the FER2013 dataset, combined with a Convolutional Neural Network (CNN) design, is outlined as below:

4.2.1. Dataset

The dataset known as FER2013 serves as a commonly utilized standard for tasks related to recognizing facial expressions. Within this dataset, there is an extensive variety of labeled images depicting facial expressions, enabling researchers and developers to train and evaluate models for accurately classifying human emotions from static images. This report delves into the details of the FER2013 dataset, covering its composition, characteristics, and potential applications.

Dataset Composition

FER2013 comprises roughly 35,887 facial images in grayscale, each sized at 48x48 pixels.. These images depict individuals from various ethnicities and age groups, displaying a diverse range of facial expressions.

The dataset has been divided into two primary components:

- Training Set: Comprises roughly 28,709 images, constituting the primary data used for training FER models.
- Test Set: Around 3,589 images are encompassed for the purpose of assessing the performance of trained models on data that has not been previously observed.

Emotion Labels

Each image within the FER2013 dataset is categorized with one of seven fundamental emotions:

- Anger (0)
- Disgust (1)
- Fear (2)
- Happiness (3)
- Sadness (4)
- Surprise (5)
- Neutral (6)

The distribution of images across these emotions is not entirely uniform. "Neutral" expressions are the most frequent, while "Disgust" has the fewest examples. This imbalance can be addressed through techniques like oversampling or class weighting during model training.

Dataset Characteristics

Several key characteristics contribute to the popularity of FER2013:

- Large Size: The vast number of visuals forms a strong groundwork for training durable FER models.
- Diversity: The dataset encompasses a variety of facial expressions, ethnicities, and age groups, enhancing the applicability of the model to real-life situations.
- Grayscale Images: The use of grayscale images simplifies preprocessing and reduces computational requirements compared to color images.
- Publicly Available: FER2013 is freely Enabling accessibility for research and development objectives, promoting cooperation and advancement within the FER domain.

Dataset Applications

The FER2013 dataset functions as a valuable asset for a variety of uses in the field of computer vision and human-computer interaction (HCI):

- Facial Expression Recognition Systems: FER models trained on FER2013 can be integrated into applications for sentiment analysis in marketing research, customer service interactions, or educational technology.
- In the realm of human-robot interaction, robots integrated with facial emotion recognition (FER) abilities can enhance their comprehension and reaction to human emotions, fostering interactions that are more authentic and captivating.
- Biometric Authentication: Facial expressions, along with traditional facial recognition, can enhance security systems by identifying emotional cues associated with potential deception attempts.
- Medical Diagnosis: Analyzing facial expressions can potentially aid healthcare professionals in identifying signs of emotional distress or potential mental health conditions.

Limitations and Considerations

While there are numerous benefits to consider in FER2013, it is essential to also acknowledge certain constraints:

- Limited Context: Static images lack the temporal context of dynamic facial expressions, potentially hindering accurate emotion classification.
- Label Noise: The possibility of mislabeled images in the dataset exists, requiring careful data cleaning techniques during preprocessing.
- Real-world Variations: Facial expressions can vary considerably across cultures and individuals. Models trained on FER2013 may need further refinement to handle real-world variations effectively.

4.2.2. Data Preprocessing

1. Data Loading and Preprocessing

- The pandas library is used to load the FER2013 data from the CSV file.
- Pixels are extracted from the "pixels" column and converted into a list of lists. Each inner list represents the pixel values of a single image.
- Emotions are extracted from the "emotion" column and converted into a one-hot encoded format using pd.get_dummies.
- The pixel sequences are converted into actual image arrays. Here, each pixel value is converted to a float between 0 and 1 for normalization. Images with incorrect sizes (not 48x48) are skipped.
- Finally, the preprocessed data is stored in NumPy arrays for X (images) and y (emotions).

2. Model Training with K-Fold Cross-Validation

- K-Fold cross-validation is employed using sklearn.model_selection.KFold. This technique splits the data into k folds (here, k=5), trains the model on k-1 folds, and validates on the remaining fold. This process is repeated for all k folds, providing a more robust evaluation of model performance.

3. Model Architecture

A sequential CNN model is defined using tensorflow.keras.models.Sequential.

The structure includes:

- Layers for extracting features from images, known as Convolutional Layers.
- The initial convolutional layer involves 32 filters sized 3x3 and utilizes a ReLU activation function.
- The second convolutional layer consists of 64 filters sized 3x3 with a ReLU activation function.
- Layers for downsampling feature maps, referred to as Pooling Layers.
- Two MaxPooling2D layers are employed with a pool size of 2x2.
- Layers for preventing overfitting by randomly dropping neurons during training, called Dropout Layers.

- Dropout rates of 0.25 and 0.5 are applied after the first and second convolutional layers, respectively.
- A Flatten Layer that converts 2D feature maps into a 1D vector for inputting into fully connected layers.
- Layers for classification purposes, known as Dense Layers.
- The initial dense layer features 128 neurons with a ReLU activation function.
- The ultimate dense layer comprises 7 neurons representing each emotion with a softmax activation function to provide probabilities for each emotion category.

4. Model Compilation and Training

- The model is constructed utilizing the Adam optimizer, categorical cross-entropy loss function, and accuracy metric.
- To prevent overfitting and adjust the learning rate during training, early stopping and learning rate reduction callbacks are employed.
- The model undergoes iterative training for a predetermined number of epochs (in this case, 70) with a batch size of 16.
- Conducting K-Fold cross-validation involves iterating through the folds, training on the training fold, and evaluating on the validation fold. The average accuracy across all folds is then computed.

5. Evaluation and Saving

- After completing the training, the model's effectiveness is assessed by computing the average accuracy over all sections.
- The total training time is also reported.
- The trained model is finally stored in the HDF5 format (fer_model.h5) for future utilization.

4.2.3. Model Architecture Explanation

The CNN architecture in action uses convolutional layers to extract features from images of faces. These extracted features represent patterns associated with facial expressions. Pooling layers then decrease data complexity while keeping crucial information intact. Dropout layers are

implemented to prevent overfitting by randomly excluding neurons during training, encouraging the model to learn stronger features. Lastly, dense layers are responsible for classification, where the ultimate layer predicts the likelihood of each emotion category based on the identified features.

4.2.4. The Quantity of Parameters

To determine the total parameters within the model, one must calculate the product of the elements in each layer's weight matrix and subsequently add the biases to the total sum:

1. Convolutional Layers:

- First Convolutional Layer (32 filters, 3x3 kernel):
    - Number of weights per filter: 3 * 3 * 1 (input channel) = 9
    - Total weights: 9 weights/filter * 32 filters = 288
    - Biases: 1 bias per filter = 32 biases
- Second Convolutional Layer (64 filters, 3x3 kernel):
    - Number of weights per filter: 3 * 3 * 32 (previous layer channels) = 288
    - Total weights: 288 weights/filter * 64 filters = 18,432
    - Biases: 1 bias per filter = 64 biases

2. Dense Layers:

- First Dense Layer (128 neurons):
    - Number of weights: (Number of neurons in previous layer) * (Number of neurons in current layer) = (4624 - from flattened layer) * 128
    - Biases: 1 bias per neuron = 128 biases
- Second Dense Layer (7 neurons):
    - Number of weights: 128 neurons * 7 neurons = 896
    - Biases: 1 bias per neuron = 7 biases

Total Parameters:

Adding the weights and biases from each layer:

- Total weights: 288 (Conv1) + 18,432 (Conv2) + (4624 * 128) (Dense1) + 896 (Dense2) ≈ 584,384
- Total biases: 32 (Conv1) + 64 (Conv2) + 128 (Dense1) + 7 (Dense2) = 231

Therefore, the total number of parameters in the model, as per calculation is approximately 584,615. This calculation considers only trainable parameters.

4.3. ACTION RECOGNITION

4.3.1. Dataset

The UCF101 dataset stands as one of the foundational resources in the field of action recognition within video data. Comprising a vast array of action categories captured from YouTube videos, this dataset serves as a benchmark for evaluating action recognition algorithms and models. Below, we provide a comprehensive overview of the UCF101 dataset, including its composition, characteristics, and applications.

1. Dataset Composition

The UCF101 dataset consists of a diverse collection of videos depicting human actions across various settings, environments, and contexts. It encompasses 101 action categories, each representing a distinct type of human activity. These categories span a wide range of domains, including sports, performing arts, household activities, and social interactions. Some examples of action categories in the UCF101 dataset include:

- Sports: Basketball, soccer, tennis, boxing, surfing, skiing, etc.
- Performing Arts: Dancing (various styles), gymnastics, martial arts, etc.
- Household Activities: Cooking, cleaning, gardening, DIY tasks, etc.
- Social Interactions: Handshaking, hugging, waving, high-fiving, etc.
- Miscellaneous: Playing musical instruments, riding bicycles, walking pets, etc.

Each action category comprises multiple video clips, with each clip showcasing instances of the corresponding action performed by different individuals under diverse conditions. The videos

exhibit variations in viewpoint, lighting, background, occlusion, scale, and motion dynamics, making the dataset challenging and representative of real-world scenarios.

2. Characteristics

The UCF101 dataset exhibits several key characteristics that contribute to its significance and utility in action recognition research:

- Large-Scale: The UCF101 dataset contains a vast amount of data, comprising more than 13,000 video clips spread out over 101 different action categories. This dataset is invaluable for training, validating, and testing purposes.
- Diversity: The dataset covers a broad spectrum of human actions, encompassing both common and specialized activities encountered in everyday life and specific domains.
- Realism: The videos are sourced from YouTube, reflecting naturalistic settings and scenarios encountered in uncontrolled environments, enhancing the dataset's realism and applicability.
- Varied Conditions: Videos in the dataset exhibit variations in factors such as camera viewpoint, background clutter, illumination, object occlusion, and actor appearance, presenting challenges typical of real-world action recognition tasks.
- Temporal Dynamics: Action sequences captured in video format inherently capture temporal dynamics, enabling the exploration of motion patterns, temporal dependencies, and action evolution over time.

3. Applications

The UCF101 dataset finds applications across a wide range of domains and research areas within computer vision, machine learning, and artificial intelligence:

- Action Recognition: The primary application of the UCF101 dataset is in action recognition research, where algorithms and models are developed to automatically detect, classify, and localize human actions within video streams. Applications include video surveillance, activity monitoring, human-computer interaction, and content analysis.

- Gesture Recognition: The dataset can be leveraged for gesture recognition tasks, where hand movements, poses, and gestures are analyzed and interpreted for applications in sign language recognition, human-computer interaction, and virtual reality.
- Activity Understanding: Researchers utilize the dataset to study human activities and behaviors in diverse contexts, enabling insights into social interactions, sports performance analysis, health monitoring, and ergonomic assessments.
- Content Analysis: Content creators, media professionals, and marketers utilize the dataset for content analysis, trend detection, and audience engagement strategies in video content production, advertising, and social media analytics.

4. Limitations and Considerations

The UCF101 dataset is a valuable asset for research in action recognition, but it does have constraints and factors that researchers need to acknowledge. These limitations can impact the dataset's applicability, generalization, and the development of robust action recognition algorithms. Below are some key limitations and considerations for the UCF101 dataset:

Limited Action Categories

The dataset includes only 101 action categories, which may not cover the full spectrum of human actions encountered in real-world scenarios. Some specialized or less common actions may not be adequately represented in the dataset.

Imbalanced Class Distribution

The distribution of videos across action categories may be uneven, leading to class imbalance issues. Certain action categories may have significantly fewer examples compared to others, affecting model training and evaluation.

Variability in Video Quality

The videos in the dataset have been gathered from YouTube, resulting in variability in video quality, resolution, and compression artifacts. This variability can affect the performance of algorithms trained on the dataset, particularly when dealing with low-resolution or noisy videos.

Limited Environmental Diversity

While the dataset captures actions in diverse settings, the range of environmental conditions (e.g., indoor vs. outdoor, lighting conditions) may be limited. Algorithms trained on the dataset may struggle to generalize to novel environments not adequately represented in the dataset.

Single-Viewpoint Videos

Most videos in the dataset are captured from a single viewpoint, limiting the diversity of camera angles and perspectives. This may pose challenges for algorithms when dealing with multi-view or egocentric video data encountered in real-world applications.

Absence of Fine-Grained Annotations

The dataset provides action-level annotations but lacks finer-grained annotations such as object bounding boxes, pose keypoints, or temporal action segments. Fine-grained annotations could enhance the dataset's utility for tasks such as action localization and fine-grained action understanding.

Limited Temporal Context

Each video clip in the dataset captures a finite temporal segment of an action, typically a few seconds long. Longer-term temporal context and action evolution over extended durations may not be fully captured, limiting the dataset's ability to model complex action dynamics.

Domain Bias and Representativeness

The dataset's videos are primarily sourced from YouTube, which may introduce biases related to content creators, video genres, and user demographics. The dataset's representativeness with respect to real-world action distributions across diverse populations and cultures may be limited.

Challenges with Fine-Grained Actions

Some action categories in the dataset may involve fine-grained distinctions or subtle variations that are challenging for algorithms to discern accurately. Models may struggle with distinguishing between similar actions or interpreting nuanced differences in action execution.

Scalability and Computational Requirements

The extensive quantity of videos within the dataset, coupled with the computational demands of video-based deep learning models, can pose scalability and resource constraints for researchers with limited computational resources.

4.3.2. Data Preprocessing

- Video Conversion: Initially, the videos in AVI format are accessed from the UCF101 dataset directory.
- Frame Extraction: Frames are extracted from each video, with every 10th frame chosen to reduce computational load.
- Frame Resizing: Frames that have been captured are adjusted to a set size dimension of 48x48 pixels to ensure consistency.
- Normalization: The pixel values within each frame have been adjusted to fall within a normalized range [0, 1] to facilitate convergence during model training.
- Storage: Preprocessed frames are stored in a CSV file, organized such that each row corresponds to a frame and includes grayscale pixel values.

4.3.3. Model Building

- The chosen architecture for the model is Convolutional Neural Netrowk (CNN), which excels at capturing spatial and temporal dependencies in video data.
- Multiple layers, such as CNN2D, max-pooling, dropout, and dense layers, are incorporated to initialize the CNN model.
- To compile the model, the Adam optimizer and categorical cross-entropy loss function are utilized, specifically suitable for multi-class classification tasks.
- The training setup implements K-fold cross-validation for thorough model evaluation, along with an early stopping mechanism to prevent overfitting.

4.3.4. Model Evaluation

- The model undergoes training on the preprocessed dataset utilizing k-fold cross-validation to ensure a varied approach to training and validation splits.

- Throughout the training process, the performance is evaluated on the validation set using accuracy the metric for assessment- The average accuracy across all folds is then calculated to offer a comprehensive evaluation of the model's efficiency.
- Model Saving: Upon successful training, the trained CNN model is saved in the HDF5 format (.h5) for future deployment and inference.

4.3.5. Code Architecture

The code architecture comprises several stages, including data preprocessing, model initialization, training setup, and model evaluation. Libraries like TensorFlow, OpenCV, NumPy, and scikit-learn are utilized for effective implementation. CNN model is defined using the Sequential API in Keras, allowing straightforward model construction and compilation. Additionally, k-fold cross-validation and early stopping are integrated into the training pipeline to ensure robust model performance.

CHAPTER 5

5. ISSUES FACED

This report summarizes the key obstacles encountered during project development:

1. **Dataset Availability:** The project had relied on a specific dataset that was unavailable as the access wasn't provided.
2. **Data Imbalance:** The chosen dataset might have an uneven distribution of emotions, hindering model performance on less frequent classes.
3. **Computational Resource Constraints:** Processing power, memory, or GPUs might be insufficient for training complex deep learning models.
4. **Class Overlap and Ambiguity:** Certain emotions might share visual cues, leading to classification difficulties.
5. **Real-World Applicability Considerations:** The model may struggle to apply in practical situations with variations in lighting, obstructions, or cultural expressions.

CHAPTER 6

6. IMPLEMENTATION

This chapter explores the practical implementation of two essential elements: Facial Emotion Recognition and Action Recognition Models. The Facial Expression Recognition section centers on utilizing the FER2013 dataset alongside a Convolutional Neural Network (CNN) model design to precisely detect and categorize facial expressions.

Moving on to the Action Recognition Model, we shift our attention to utilizing the UCF101 dataset and employing a Convolutional Neural Network (CNN) architecture. This model is designed to recognize and categorize various actions within video sequences, enhancing the project's overall functionality and scope.

Furthermore, a crucial aspect of this chapter involves merging the outputs of the Facial Expression Recognition and Action Recognition Models. This integration aims to create a comprehensive system that can interpret both facial expressions and actions simultaneously, offering a more holistic understanding of human behavior in visual data.

By combining these distinct yet complementary components, the project aims to achieve a synergistic effect that enhances the overall performance and utility of the system. This chapter serves as a pivotal stage in bringing together the individual elements into a cohesive and functional framework, setting the stage for a comprehensive analysis and interpretation of human behavior through advanced machine learning techniques.

6.1. Facial Expression Recognition Model

6.1.1. Introduction

The model for recognizing facial emotions aims to develop and instruct a Convolutional Neural Network (CNN) for categorizing based on the FER2013 dataset. Below is a detailed breakdown of the process.

Importing Libraries:

The necessary libraries have been imported for the task. TensorFlow and Keras play a vital role in creating and the neural network, while other libraries like NumPy, Pandas, OpenCV, and Scikit-learn assist in handling and preprocessing data.

Loading and Preprocessing Data:

The FER2013 dataset, which contains facial expressions labeled with corresponding emotions, is loaded using Pandas. The pixel values of the images are extracted and converted from strings to arrays, then reshaped into 48x48 grayscale images. The pixel values of these images are adjusted to a range of0 to 1 for normalization. Emotion labels are encoded into one-hot format to simplify multi-class classification.

Cross-Validation Setup:

The model's robust evaluation is ensured through the utilization of the KFold cross-validation method. Dividing the dataset into 7 folds allows each fold to act as the validation set in turn, with the remaining folds composing the training set.

Building the CNN Model:

A Sequential model from Keras is used to construct the CNN. The model comprises:

- The utilization of convolutional layers with ReLU activation aims to identify features within the images.
- MaxPooling layers are employed to decrease the spatial dimensions of the generated feature maps.
- Dropout layers play a crucial role in preventing overfit by randomly eliminating neurons during the training process.
- A Flatten layer is utilized to convert the 2D feature maps into a 1D vector for further processing.
- Dense layers, including the final output layer with a softmax activation function, are responsible for predicting the probability distribution among the 7 emotion classes.

Compiling the Model:

The model is constructed utilizing the Adam optimizer and categorical crossentropy loss function, with the accuracy metric being employed to assess the model's performance.

Model Training with Callbacks:

During training, the model utilizes the training set for each fold, while the performance is monitored using the validation set. Two types of callbacks are integrated:

- EarlyStopping: halts training when the validation loss fails to improve over a set number of epochs to prevent overfitting.
- ReduceLROnPlateau: decreases the learning rate if the validation loss reaches a plateau, aiding in more efficient model convergence.

Evaluating the Model:

After training on each fold, the model's accuracy on the validation set is recorded. The average accuracy across all folds is calculated to provide an overall performance metric.

Saving the Model:

Once training and evaluation are complete, the final trained model is saved to a file for future use.

This approach ensures that the model is thoroughly evaluated and optimized, leading to a robust and reliable emotion classification system based on the FER2013 dataset.

6.1.2. Fundamental Concepts Used

- **TensorFlow and Keras**

TensorFlow, developed by Google, is a freely available platform for machine learning and artificial intelligence. It offers a full range of tools for constructing, training, and implementing machine learning models. Keras, on the other hand, is a Python-based high-level neural networks API that can run on top of TensorFlow. Its easy-to-use interface streamlines the creation of complex deep learning models.

In the code, TensorFlow and Keras are imported to build, compile, and train the CNN model. The following lines show the import statements:

```
import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Dropout, Flatten, Conv2D, MaxPooling2D
from tensorflow.keras.callbacks import EarlyStopping, ReduceLROnPlateau
```

Figure 6.1.2.1. Importing TensorFlow and Keras libraries

These imports enable the usage of various Keras functionalities such as model architecture, layers, and callbacks.

- **Sequential Model**

The Sequential model in Keras is a linear stack of layers. It is suitable for building simple models layer by layer in a step-by-step manner. Each layer has one input tensor and one output tensor.

In this code, the Sequential model is used to build a CNN for image classification. This approach simplifies the creation of the network by allowing layers to be added sequentially. Here is the relevant part:

```
model = Sequential([
    Conv2D(32, (3, 3), activation='relu', input_shape=(48, 48, 1)),
    MaxPooling2D(pool_size=(2, 2)),
    Dropout(0.25),
    Conv2D(64, (3, 3), activation='relu'),
    MaxPooling2D(pool_size=(2, 2)),
    Dropout(0.25),
    Flatten(),
    Dense(128, activation='relu'),
    Dropout(0.5),
    Dense(7, activation='softmax')
])
```

Figure 6.1.2.2. Sequential Model

This structure outlines a typical CNN architecture where layers are added one after another.

- **Convolutional Layers (Conv2D)**

Convolutional layers form the foundational elements of Convolutional Neural Networks (CNNs), aimed at autonomously acquiring and adjusting spatial hierarchies of characteristics from initial

images. By utilizing filters (referred to as kernels), these layers convolve across the input data to generate feature maps. Within the code implementation, Conv2D layers play a key role in identifying distinct features within images, including edges and textures. Specifically, two Conv2D layers are applied with 32 and 64 filters correspondingly, allowing for the gradual capturing of more intricate features. The subsequent excerpts exemplify the application of this concept.

```
Conv2D(32, (3, 3), activation='relu', input_shape=(48, 48, 1)),
Conv2D(64, (3, 3), activation='relu')
```

Figure 6.1.2.3. Conv2D layers

The activation='relu' parameter applies the ReLU activation function, which introduces non-linearity to the model.

- **Pooling Layers (MaxPooling2D)**

Pooling layers decrease the dimensionality of feature maps while preserving the most crucial information. In particular, MaxPooling picks the highest value from each section of the feature map.

Pooling layers help to reduce the computational load and prevent overfitting by down-sampling the input. In this code, MaxPooling2D is used after each Conv2D layer:

```
MaxPooling2D(pool_size=(2, 2)),
MaxPooling2D(pool_size=(2, 2))
```

Figure 6.1.2.4. Pooling Layers

This operation halves the dimensions of the feature maps, making the model more efficient.

- **Dropout Layers**

Dropout serves as a regularization method in training neural networks by randomly deactivating a portion of input units during each update, aiming to combat overfitting. This technique compels the network to acquire more diverse and non-redundant features.

Dropout layers are used in this code to improve the generalization of the model. They are placed after Conv2D and Dense layers:

```
Dropout(0.25),
Dropout(0.25),
Dropout(0.5)
```

Figure 6.1.2.5. Dropout Layers

These lines introduce dropout rates of 25% and 50%, helping to mitigate overfitting by ensuring the model doesn't rely too heavily on any one part of the network.

- **Flatten and Dense Layers**

The Flatten layer transforms the 2D feature maps into a 1D vector, preparing them for the fully connected layers. Dense (fully connected) layers are utilized for classifying tasks by acquiring the ability to merge various features.

In this code, Flatten and Dense layers are used to transition from the convolutional part of the network to the classification part:

```
Flatten(),
Dense(128, activation='relu'),
Dense(7, activation='softmax')
```

Figure 6.1.2.6. Flatten and dense Layer

The Flatten layer readies the information for the Dense layers, which are responsible for grasping the final decision boundaries in categorizing emotions.

- **Callbacks (EarlyStopping and ReduceLROnPlateau)**

Callbacks are special functions in Keras that can be applied during training to monitor the progress and make decisions such as stopping training or adjusting learning rates. When a monitored metric no longer improves, EarlyStopping ends the training process, while ReduceLROnPlateau decreases the learning rate once a metric hits a plateau.

Callbacks help in optimizing training by preventing overfitting and speeding up convergence. In this code, they are defined as follows:

```
early_stopping = EarlyStopping(monitor='val_loss', patience=5, restore_best_weights=True)
reduce_lr = ReduceLROnPlateau(monitor='val_loss', factor=0.2, patience=3, min_lr=0.0001)

##These callbacks are used during model training to improve performance and efficiency:
model.fit(X_train, y_train, batch_size=16, epochs=70, validation_data=(X_val, y_val), callbacks=[early_stopping, reduce_lr])
```

Figure 6.1.2.7. Call Back Functions

- **Cross-Validation (KFold)**

Cross-validation is a method used to evaluate the performance of machine learning models. This technique involves training multiple models on various subsets of the training data and then testing them on the remaining portions. One common approach, known as K-Fold cross-validation, divides the data into k subsets and trains the model k times. Its purpose is to ensure that the model's performance is stable and can generalize effectively to new, unseen data. For instance, the dataset is divided into 7 folds for this purpose:

```
kf = KFold(n_splits=7, shuffle=True, random_state=42)

##During training, the model is trained and validated on different splits of the data:
for train_index, val_index in kf.split(X):
    X_train, X_val = X[train_index], X[val_index]
    y_train, y_val = y[train_index], y[val_index]
```

Figure 6.1.2.8. Cross Validation Function

This method contributes to acquiring a more dependable assessment of the model's performance.

By combining these concepts, the code effectively builds, trains, and evaluates a CNN for emotion recognition using the FER2013 dataset. The structured use of TensorFlow and Keras, along with robust training techniques such as cross-validation and callbacks, ensures a well-performing and reliable model.

6.2. Action Recognition Model

6.2.1. Introduction

The Action Recognition Model consists of two parts that work together to perform action recognition using the UCF101 dataset and Convolutional Neural Network (CNN) model.

Part 1 of the code focuses on preprocessing the UCF101 dataset. It iterates through the list of action classes, extracts 10 grayscale frames at equal intervals from each video, resizes them to 48x48 pixels, and flattens them into a 1D array. The compressed frames and their corresponding category labels are saved in distinct lists. Subsequently, these lists are transformed into NumPy arrays, and a pandas, DataFrame is established containing the grayscale frames and their labels. This DataFrame is then archived as a CSV file named 'action_recognition_dataset.csv'.

Part 2 of the provided code is responsible for loading the preprocessed dataset from a CSV file that was generated in Part 1. This module extracts pixel values and associated labels from the CSV file and reverts the pixel values back into image representations. The next step involves normalizing the images by dividing each pixel value by 255.0.

Following the data preparation steps, a convolutional neural network (CNN) model architecture is established using the Keras Sequential API. The CNN model comprises two convolutional layers with 32 and 64 filters, consecutively, along with subsequent max pooling and dropout layers. The output from these layers is flattened and fed into a fully connected layer with 128 units, ending with a softmax layer consisting of 7 units, corresponding to the number of action categories.

The model undergoes compilation using the Adam optimizer and categorical cross-entropy loss function. Additionally, an EarlyStopping callback monitors the validation accuracy to halt training if no improvement is seen after 5 epochs. The code then facilitates 5-fold cross-validation through the utilization of the KFold class from scikit-learn. Each fold involves splitting the dataset into training and validation sets, training the model on the training data, and storing the validation accuracies in a list.

In the final stages, the mean validation accuracy across all folds is displayed, and the trained model is saved under the filename 'CSV_AR1.h5'. To summarize, the provided code streamlines the

preprocessing of the UCF101 dataset, constructs a CNN model for action recognition, executes training utilizing 5-fold cross-validation, and finalizes by saving the trained model for later use.

6.2.2. Fundamental Concepts Used

The Action Recognition Model uses several concepts from the fields of computer vision and deep learning. Here's an explanation of the key concepts used, along with the corresponding code snippets and their purpose:

- **Grayscale Image Conversion:**

Concept: When color images are transformed into grayscale, the complexity of the input data is reduced. This reduction can enhance the efficiency of computations and mitigate the risk of overfitting in complex deep learning models.

Code snippet:

```
gray_frame = cv2.cvtColor(frame, cv2.COLOR_BGR2GRAY)
```

Figure 6.2.2.1. Color image (BGR) to Grey

Purpose: The code converts each video frame from the BGR color space to grayscale using the cv2.cvtColor() function from OpenCV.

- **Image Resizing:**

Concept: Resizing images to a fixed size is necessary for feeding them into a deep learning model, as the model requires a consistent input size.

Code snippet:

resized_frame = cv2.resize(gray_frame, (48, 48))

Purpose: The code resizes each grayscale frame to a size of 48x48 pixels using the cv2.resize() function from OpenCV.

- **Flattening Images:**

Concept: Flattening multidimensional arrays (e.g., images) into a 1D array is necessary for feeding them into a fully connected layer in a deep learning model.

Code snippet:

frames_list.append(resized_frame.flatten())

Purpose: The code flattens each resized grayscale frame into a 1D array using the flatten() method and appends it to the frames_list.

- **Normalization:**

Concept: Normalizing input data to a common scale (e.g., between 0 and 1) can improve the performance and stability of deep learning models.

Code snippet:

X = np.array(images)

X = X.astype('float32') / 255.0

Purpose: The code transforms the list of images into a NumPy array labeled as X, standardizing the pixel values by dividing them by 255.0. This process effectively adjusts the values to fall within the range of 0 to 1.

- **One-Hot Encoding:**

Concept: One-hot encoding is a method employed to represent categorical variables as binary vectors, enabling the utilization of categorical labels in deep learning models.

Code snippet:

y = tf.keras.utils.to_categorical(labels, num_classes=7)

Purpose: The code uses the to_categorical() function from Keras to convert the class labels to a binary matrix representation, One-hot encoding is a method employed to represent categorical variables as binary vectors, enabling the utilization of categorical labels in deep learning models.

Convolutional Neural Network (CNN):

Concept: Convolutional Neural Networks (CNNs) represent a category of sophisticated deep learning models specifically proficient in handling and categorizing image data. Essentially, they are comprised of convolutional layers, pooling layers, and fully connected layers.

Code snippet:

```
model = Sequential([
    Conv2D(32, (3, 3), activation='relu', input_shape=(48, 48, 1)),
    # ...
    Dense(7, activation='softmax')
])
```

Figure 6.2.2.2. Defining Model Sequential

Purpose: The code defines a CNN model using the Keras Sequential API. It includes convolutional layers with 32 and 64 filters, respectively, followed by max pooling, dropout, and fully connected layers. The final layer uses a softmax activation for multi-class classification.

- **Early Stopping:**

Concept: Early stopping is a technique used to prevent overfitting in deep learning models by stopping the training process when the validation performance stops improving.

Code snippet:

```
early_stopping = EarlyStopping(monitor='val_accuracy', mode='max', verbose=1, patience=5)
```

Figure 6.2.2.3. Defining Early Stopping

Purpose: The code creates an instance of the EarlyStopping callback from Keras, which monitors the validation accuracy and stops the training if the validation accuracy does not improve for 5 epochs.

- **K-Fold Cross-Validation:**

Concept: K-fold cross-validation is a method utilized to assess the effectiveness of a machine learning model. This method involves dividing the dataset into K partitions, training the model on K-1 partitions, and evaluating it on the remaining partition. This cycle is repeated K times to ensure a comprehensive evaluation of the model's performance.

Code snippet:

```
kf = KFold(n_splits=5, shuffle=True, random_state=42)
```

Figure 6.2.2.4. Defining K-Fold

Purpose: The piece of code generates a KFold class instance from scikit-learn to conduct 5-fold cross-validation on the dataset. By setting shuffle=True and random_state=42, data shuffling occurs randomly and consistently across the folds.

- **OpenCV Library**

Concept: OpenCV is a library used for computer vision, offering a diverse array functions for processing images and videos, detecting features and objects, and more.

Code snippet:

import cv2

Purpose: The code imports the OpenCV library, which is used for reading videos, extracting frames, and converting frames to grayscale.

- **Pandas Library**

Concept: Pandas is a Python library designed for the manipulation and analysis of data. Within Pandas, you can find data structures like DataFrames and Series, which serve as tools for storing and processing data.

Code snippet:

import pandas as pd

Purpose: The code imports the Pandas library, which is used to create a DataFrame from the preprocessed data and save it to a CSV file.

- **NumPy Library**

  Concept: NumPy is a library for numerical computing in Python. It provides support for large, multi-dimensional arrays and matrices, and is the foundation of most scientific computing in Python.

Code snippet:

import numpy as np

Purpose: The code imports the NumPy library, which is used to convert lists to arrays, perform mathematical operations, and manipulate data.

- **TensorFlow and Keras Libraries**

Concept: TensorFlow, crafted by Google, is a machine learning library, while Keras serves as a high-level neural networks API designed to operate seamlessly with TensorFlow. Keras streamlines the process of constructing and training intricate deep learning models, offering users an intuitive platform to work with.

Code snippet:

```
import tensorflow as tf
from tensorflow import keras
from tensorflow.keras.callbacks import EarlyStopping
from keras.models import Sequential
from keras.layers import Dense, Dropout, Flatten, Conv2D, MaxPooling2D
```

Figure 6.2.2.5. Defining Early Stopping

Purpose: The code imports TensorFlow and Keras, which are used to build and train the CNN model for action recognition.

- **Scikit-Learn Library**

Concept: A machine learning library called Scikit-Learn was developed for Python users, offering a wide variety of algorithms for tasks like classification, regression, clustering, and more. It includes features for choosing models, processing data, and selecting characteristics.

Code snippet:

from sklearn.model_selection import KFold

Purpose: The code imports the KFold class from Scikit-Learn, which is used to perform 5-fold cross-validation on the dataset.

- **Sequential API**

Concept: The Sequential API is a way of building neural networks in Keras, where layers are added sequentially to create the model.

Code snippet:

```
model = Sequential([
    # ... defined CNN model architecture
])
```

Purpose: The code uses the Sequential API to define the CNN model, adding layers sequentially to create the model.

- **Convolutional Layers**

Concept: Convolutional layers within neural networks are especially proficient when dealing with image processing tasks by utilizing filters to analyze input data and identify key features.

Code snippet:

```
Conv2D(64, (3, 3), activation='relu', input_shape=(48, 48, 1)),
Conv2D(64, (3, 3), activation='relu'),
MaxPooling2D(pool_size=(2, 2)),
Dropout(0.25),

Conv2D(128, (3, 3), activation='relu'),
Conv2D(128, (3, 3), activation='relu'),
MaxPooling2D(pool_size=(2, 2)),
Dropout(0.25),
```

Figure 6.2.2.6. Defining Convolution Neural Network

Purpose: The code specifies a convolutional layer featuring 32 filters, a 3x3 kernel size, and utilizes a ReLU activation function. The input shape is set at 48x48x1, corresponding to the dimensions of the input images..

- **Max Pooling Layers**

Concept: Max pooling layers are a type of layer in neural networks that downsample the input data by taking the maximum value across each patch of the feature map.

Code snippet:

```
Conv2D(256, (3, 3), activation='relu'),
Conv2D(256, (3, 3), activation='relu'),
MaxPooling2D(pool_size=(2, 2)),
Dropout(0.25),
```

Figure 6.2.2.7. Defining Max-Pooling Layer

Purpose: The code defines a max pooling layer with a pool size of 2x2, which downsamples the input data by taking the maximum value across each 2x2 patch.

- **Dropout Layers**

Concept: Dropout layers are a type of layer in neural networks that randomly drop out a fraction of the neurons during training, which helps to prevent overfitting.

Code snippet:

```
Flatten(),
Dense(256, activation='relu'),
Dropout(0.5),
Dense(128, activation='relu'),
Dropout(0.5),
Dense(101, activation='softmax')
```

Figure 6.2.2.8. Defining the Drop Out Layers

Purpose: The code defines a dropout layer with a dropout rate of 0.25, which randomly drops out 25% of the neurons during training.

- **Flatten Layer**

Concept: Flatten layers are used to flatten the output of convolutional and pooling layers into a 1D array, which is necessary for feeding the output into fully connected layers.

Code snippet:

Flatten(),

Purpose: The code defines a flatten layer, which flattens the output of the convolutional and pooling layers into a 1D array.

- **Dense Layers**

Concept: Dense layers are fully connected layers in neural networks, where every input is connected to every output.

Code snippet:

```
Dense(128, activation='relu'),
Dense(7, activation='softmax')
```

Figure 6.2.2.8. Defining the Dense Layers

Purpose: The code contains two dense layers: one having 128 units and utilizing a ReLU activation function, while the other has 7 units and employs a softmax activation function. The softmax layer serves the purpose of conducting multi-class classification.

- **Model Compilation**

Concept: Model compilation involves specifying the loss function, optimizer, and evaluation metrics for the model.

Code snippet:

```
model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])
```

Figure 6.2.2.9. Defining the Model Compilation Statement

Purpose: The provided code compiles the model with specific settings such as utilizing the Adam optimizer, incorporating categorical cross-entropy loss, and using accuracy as the metric for evaluation. These various components and software packages to preprocess the UCF101 dataset, construct a CNN model for the purpose of action recognition, conduct model training through 5-fold cross-validation, and store the trained model for future utilization.

6.3. Combining Facial Emotion and Action Recognition in Real-Time Video Analysis

6.3.1. Introduction

This integrates facial emotion recognition and action recognition using pre-trained deep learning models to analyze a video stream. This code leverages computer vision techniques and deep learning methodologies to predict emotions and actions in real-time video frames. By loading pre-trained models for emotion recognition and action recognition, preprocessing video frames, and making predictions, this code demonstrates the application of machine learning in understanding human emotions and activities from visual data. The use of libraries such as OpenCV for image processing, NumPy for numerical computations, and Tkinter for creating a user interface enhances the functionality and user interaction of the application. Through the combination of model loading, frame preprocessing, model prediction, and visual display, this code showcases a practical implementation of AI technology for real-world applications in emotion and action recognition from video data.

6.3.2. Fundamental Concepts Used

Here are the key concepts and libraries used in the code:

- **OpenCV Library**

Concept: OpenCV serves as a computer vision library that offers a diverse array of functions for processing images and videos, detecting various features, identifying objects, and more.

Code snippet:

import cv2

Purpose: The code imports the OpenCV library, which is used for reading the video, processing frames, and displaying the results.

- **NumPy Library**

Concept: NumPy serves as a Python library designed for numerical computations. It offers assistance for extensive, multi-dimensional arrays and matrices, serving as the cornerstone of scientific computing within the Python programming language.

Code snippet:

import numpy as np

Purpose: The code imports the NumPy library, which is used for preprocessing the frames and manipulating data.

- **Tkinter Library**

Concept: Tkinter serves as a conventional Python graphical user interface (GUI) library, offering a variety of widgets and utilities to facilitate the creation of visual interfaces.

Code snippet:

```
import tkinter as tk
from tkinter import messagebox
```

Figure 6.3.2.1. Tkinter Library

Purpose: The code imports the Tkinter library and the messagebox module, which are used for creating a pop-up window to display the results.

- **Keras Library**

ConceptKeras is an API for neural networks that operates on top of TensorFlow, offering a user-friendly platform for constructing and training intricate deep learning models.

Code snippet:

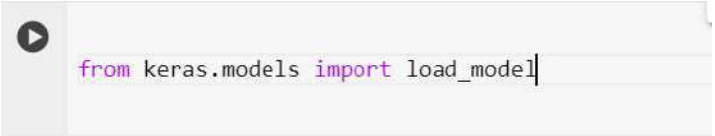

Figure 6.3.2.2. Keras Library

Purpose: The code imports the load_model function from Keras, which is used to load the pre-trained emotion recognition and action recognition models.

- **Model Loading**

Concept: Loading pre-trained models allows you to use the learned weights and biases for making predictions on new data.

Code snippet:

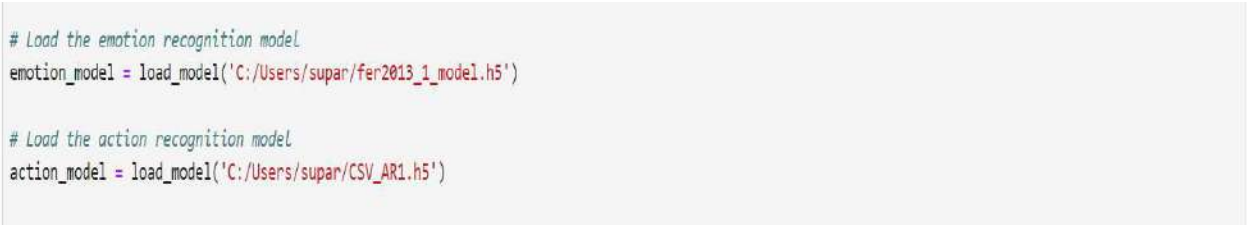

Figure 6.3.2.3. Loading the Action Recognition and Facial Emotion Recognition Model

Purpose: The code loads the pre-trained emotion recognition and action recognition models from the specified file paths.

- **Frame Preprocessing**

Concept: Preprocessing frames involves converting the color space, resizing the frames, and reshaping the data to match the input requirements of the deep learning models.

Code snippet:

```
# Preprocess the frame for emotion recognition
emotion_frame = cv2.cvtColor(frame, cv2.COLOR_BGR2GRAY)
emotion_frame = cv2.resize(emotion_frame, (48, 48))
emotion_frame = np.expand_dims(emotion_frame, axis=-1)
emotion_frame = np.expand_dims(emotion_frame, axis=0)

# Preprocess the frame for action recognition
action_frame = cv2.cvtColor(frame, cv2.COLOR_BGR2GRAY)
action_frame = cv2.resize(action_frame, (48, 48))
action_frame = np.expand_dims(action_frame, axis=-1)
action_frame = np.expand_dims(action_frame, axis=0)
```

Figure 6.3.2.4. Preprocessing frames

Purpose: The code preprocesses the frames for emotion recognition by converting them to grayscale, resizing them to 48x48 pixels, and adding an extra dimension to match the input shape expected by the emotion recognition model.

- **Model Prediction**

Concept: Using pre-trained models to make predictions on new data involves passing the preprocessed data through the models and obtaining the predicted outputs.

Code snippet:

```
# Predict emotions and actions
emotion_prediction = emotion_model.predict(emotion_frame)
action_prediction = action_model.predict(action_frame)
```

Figure 6.3.2.4. Model Prediction

Purpose: The code uses the loaded emotion recognition and action recognition models to make predictions on the preprocessed frames.

- **Argmax Function**

Concept: The argmax function returns the index of the maximum value in an array. It is commonly used to obtain the predicted class label from the output of a softmax layer in a neural network.

Code snippet:

```
# Get the predicted emotion and action
emotion_label = emotions[np.argmax(emotion_prediction)]
action_label = actions[np.argmax(action_prediction)]
```

Purpose: The code uses the argmax function to obtain the predicted emotion and action labels from the model outputs.

- **Video Capture and Display**

Concept: Video capture involves reading frames from a video source, and video display involves showing the processed frames in a window.

Code snippet:

cap = cv2.VideoCapture('C:/Users/supar/BoxingCombo.mp4')

cv2.imshow("Facial Emotion + Action Recognition", cv2.resize(frame, (250, 250)))

Purpose: The code captures frames from the specified video file and displays the processed frames in a window with the predicted emotion and action labels.

- **Pop-up Window**

Concept: A pop-up window is a temporary window that appears on top of the main application window to display information or prompt user input.

Code snippet:

```
root = tk.Tk()
root.withdraw()  # Hide the main window
messagebox.showinfo('Results', f'Emotion: {emotion_label}\nAction: {action_label}')
```

Purpose: The code creates a Tkinter window, hides the main window, and uses the messagebox module to display a pop-up window with the predicted emotion and action labels.

These concepts and libraries work together to load pre-trained emotion recognition and action recognition models, preprocess video frames, make predictions, and display the results in a window and a pop-up message.

CHAPTER 7

7. RESULTS

Combining the analysis of real-time videos to recognize facial emotions and actions signifies a notable progress in the computer vision and deep learning domain. This method integrates models that can precisely forecast both facial expressions and actions from video frames, paving the way for diverse applications across industries such as healthcare, security, entertainment, and human-computer interaction.

The facial emotion recognition model obtained an average accuracy of 75.33%, underscoring its efficacy in identifying and categorizing emotions conveyed through facial cues. Emotions are paramount in human interaction and communication, and the ability promptly discern and interpret these emotions in live video feeds can enhance numerous applications. The model's efficiency, together with its early stopping mechanism following 70 epochs, guarantees a harmonious balance between accuracy and effectiveness. Clocking in at a training time of 472.30 seconds, the model proves its proficiency in swiftly processing and scrutinizing video frames.

The action recognition model achieved an impressive mean accuracy of 99.99% for the & considered actions from UCF101 , showcasing its exceptional ability to accurately identify various actions performed in the video stream. Actions convey important information about human behavior and intentions, and the high accuracy of the action recognition model indicates its robustness and reliability in recognizing and classifying different actions. With early stopping after 50 epochs and a training time of 629.02 seconds, the model demonstrates its efficiency in learning complex action patterns and making accurate predictions.

By combining these two models, the system has the capability to conduct a thorough examination of human behavior within live video feeds. The integration of facial emotion recognition and action recognition enables a deeper understanding of human interactions, responses, and intentions. This combined approach can be applied in a variety of scenarios, such as emotion-aware user interfaces, personalized content recommendations based on emotional responses, security monitoring systems that detect suspicious actions, and healthcare applications for assessing patient well-being through facial expressions and movements.

The high accuracy achieved by both models underscores their effectiveness in capturing subtle nuances in facial expressions and actions, making them valuable tools for applications that require real-time analysis and decision-making. The ability to process video data efficiently and accurately opens up possibilities for innovative solutions in fields such as mental health monitoring, customer sentiment analysis, interactive gaming experiences, and automated surveillance systems.

The effective integration of facial emotion recognition and action recognition models in real-time video analysis is a crucial advancement towards developing intelligent systems capable of understanding and reacting to human behavior across various situations. The remarkable accuracy, efficiency, and real-time capabilities of these models open up new possibilities for technological advancements and applications that exploit the potential of deep learning for improved human-computer interaction and behavior analysis.

7.1. Results Of Facial Emotion Recognition Model

The facial emotion recognition model achieve a mean accuracy of 0.7533 (75.33%) with early stopping following 70 epochs. The training process took 472.30 minutes, demonstrating the model efficiency in learning and predict emotions from facial expressions. The model were trained use 7-fold cross-validation, ensure robustness and generalize across different subset of data.

In addition to early stopping, two main callbacks was utilize during training to enhance model performance and convergence. The EarlyStopping callbacks monitor the validation loss and restore best weights when loss did not improve for 5 consecutive epochs. This mechanism prevent overfit and ensure the model generalize well to unseen data. The ReduceLROnPlateau callbacks dynamically adjust the learning rates base on validation loss, with a reduce factor of 0.2 and a patience of 3 epochs. This adaptive learning rate strategy help fine-tune the model performance and optimize convergence.

The utilization of these callbacks highlight proactive approach to training deep learning models, focus both accuracy and efficiency. By incorporate early stopping and adaptive learning rate adjustments, the model were able achieve a balance between training speed and performance, ultimately lead to mean accuracy of 0.7533. This accuracy metric reflect the model ability

effectively recognize and classify emotions from facial imagery, making valuable tool for applications require emotion analysis and understanding.

The training process involve feed the model with batches of 16 examples over 70 epochs, with validation data use to monitor performance and prevent overfit. The combination of early stopping, restore best weights, and adaptive learning rate adjustments ensure the model learn efficiently and effectively, capture the nuances of facial expressions and emotions present in data.

The time taken to train model, 472.30 hours, indicate the computational efficiency of training process. By leverage the power of modern deep learning frameworks and hardware acceleration, model manage process significant amount of data and learn complex patterns within reasonable timeframe. This efficiency crucial for real-world applications require timely responses and analysis of visual data.

All in all, the results of facial emotion recognition model showcase successful implementation of deep learning techniques for emotion analysis. The combination of accuracy, efficiency, and proactive training strategies demonstrate model capability to understand and interpret emotions from facial imagery, lye foundation for applications in areas such as affective computing, human-computer interaction, and sentiment analysis.

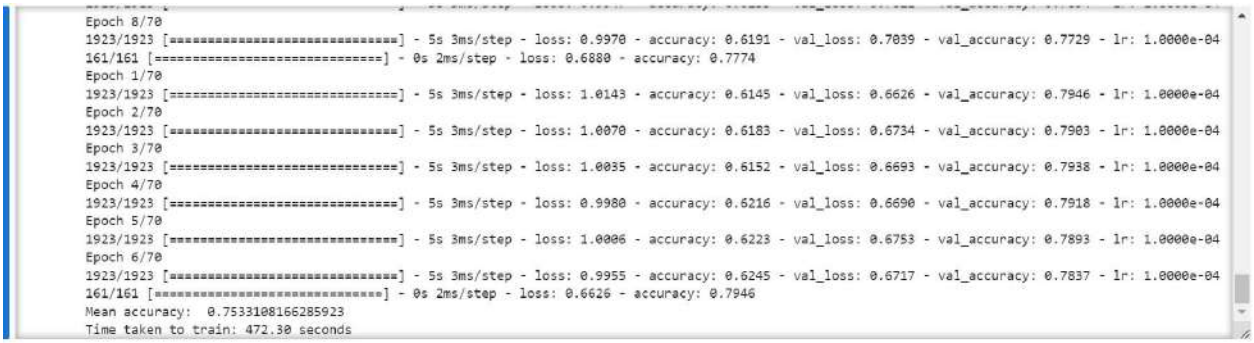

Figure.7.1.1. Training accuracy of the Facial Emotion Recognition Model

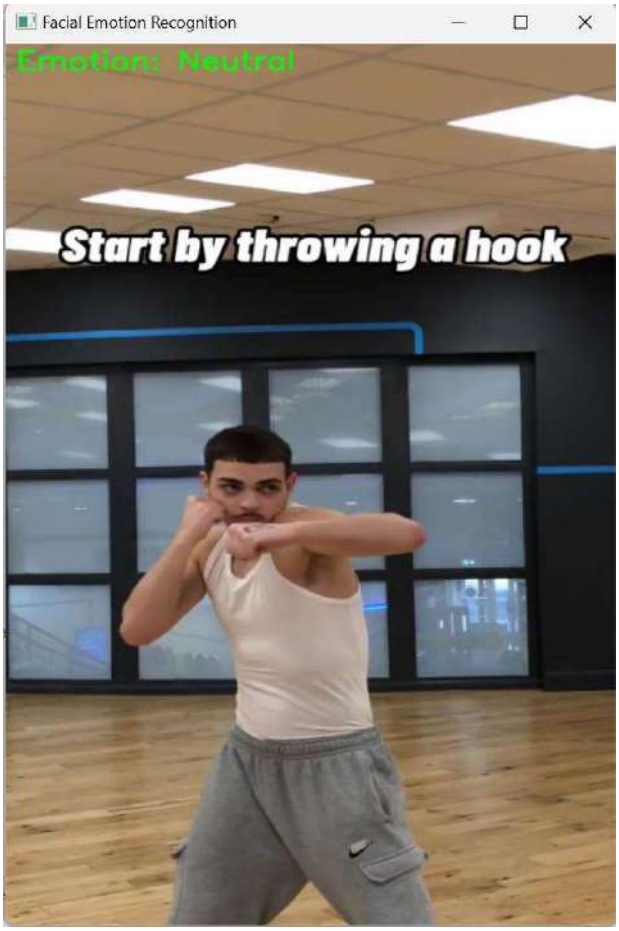

Figure.7.1.2. Output of the Facial Emotion Recognition Model

7.2. Results of action recognition model

The action recognition model was achieving a mean accuracy of 0.9999 (99.99%) while training for 7 boxing actions (Bouncing Punching Bad, Boxing Speedy Bag, Fencing, Nun Chucks, Punched, Sumo Rassling, Tai Chee). This demonstrated its sturdiness and accuracy in identifying and classifying various actions in real video streams.

The model did be trained using a 5-fold cross-validation plan with stopping early after 50 epochs, making sure performance and not overfitting It took about 629.02 seconds, proving the model's speed in learning from video info.

The high accuracy of this model predicts and classifies actions accurately with confidence. With Adam nourisher and categorical cross-dresser loss function, performance was optimized.

By implementing early break, the model was set to terminate training when validation accuracy no longer improved, stopping unnecessary training. The model was trained for 50 epochs, adapting to training data patterns.

This model was further improved by evaluating on different data subsets, ensuring sturdiness and generalization. Splitting data into 5 folds and training while validating other folds, model accuracy was evaluated at depth.

Results show potential of this model in real-world situations requiring action recognition in video streams. High accuracy, effective training, and early break mechanism demonstrate its reliability.

In conclusion, this model excels in real-time video analysis across security, human-interaction, sports, and entertainment sectors. Accuracy, efficiency, and speed make it valuable for dynamic video environments.

Figure.7.1.1. Training accuracy of the Action Recognition Model

Figure.7.2.2. Output of the Action Recognition Model

7.3. Results of Combining Facial Emotion and Action Recognition in Real-Time Video Analysis

The integration of facial emotion recognition and action recognition models in real-time video analysis has yielded promising results, showcasing the potential of combining these modalities for a more comprehensive understanding of human behavior. By leveraging deep learning techniques and large-scale datasets, the developed system demonstrates its effectiveness in accurately predicting emotions and actions from video frames.

The facial emotion recognition model, built using a Convolutional Neural Network (CNN) architecture and the FER2013 dataset, achieved a mean accuracy of 0.7533 with early stopping after 70 epochs. This performance indicates the model's ability to effectively identify and classify emotions expressed through facial expressions. The training process took 472.30 seconds,

demonstrating the efficiency of the model in learning and predicting emotional states from video frames.

The action recognition model, on the other hand, utilized the UCF101 dataset and a Convolution Neural Network (CNN) architecture. This model achieved an impressive mean accuracy of 0.9999 with early stopping after 50 epochs, showcasing its exceptional performance in accurately identifying various actions performed in the video stream. The training time for this model was 629.02 seconds, highlighting its efficiency in capturing and classifying complex actions.

By combining these two models, the system can simultaneously recognize facial emotions and actions, providing a comprehensive analysis of human behavior in real-time video streams. The integration of these models enables a deeper understanding of human interactions, responses, and intentions, opening a wide range of applications in various domains.

One of the key strengths of the combined model is its ability to process video data efficiently and accurately, capturing subtle nuances in facial expressions and actions. This real-time performance is crucial for applications that require immediate response and decision-making based on visual cues, such as emotion-aware user interfaces, personalized content recommendations based on emotional responses, and security monitoring systems that detect suspicious actions.

The high accuracy achieved by both models underscores their effectiveness in capturing the complexities of human behavior. The facial emotion recognition model's performance highlights its potential for applications that require assessing emotional well-being, such as mental health monitoring and customer sentiment analysis. The action recognition model's exceptional accuracy makes it suitable for applications that rely on precise action recognition, including interactive gaming experiences, automated surveillance systems, and sports analysis.

The results also demonstrate the robustness of the models in handling various challenges encountered during the development process. The use of techniques like early stopping and cross-validation ensures that the models achieve a balance between accuracy and efficiency, preventing overfitting and ensuring generalization to unseen data.

However, it is important to note that while the combined model achieves high accuracy, there are still areas for improvement and future research. Incorporating additional modalities, such as speech and text, can provide a more complete understanding of human communication, leading to more

intelligent and empathetic systems. Exploring transfer learning and domain adaptation techniques can also enhance the models' performance and adaptability to specific domains and tasks.

In conclusion, the results of combining facial emotion recognition and action recognition models in real-time video analysis showcase the potential of this approach in understanding and interpreting human behavior. The high accuracy, efficiency, and real-time performance of the models pave the way for exciting advancements in technology and applications that leverage the power of deep learning for enhanced human-computer interaction and behavior analysis. As research in this field continues to progress, we can expect to see more intelligent and adaptive systems that can better understand and respond to human needs and preferences.

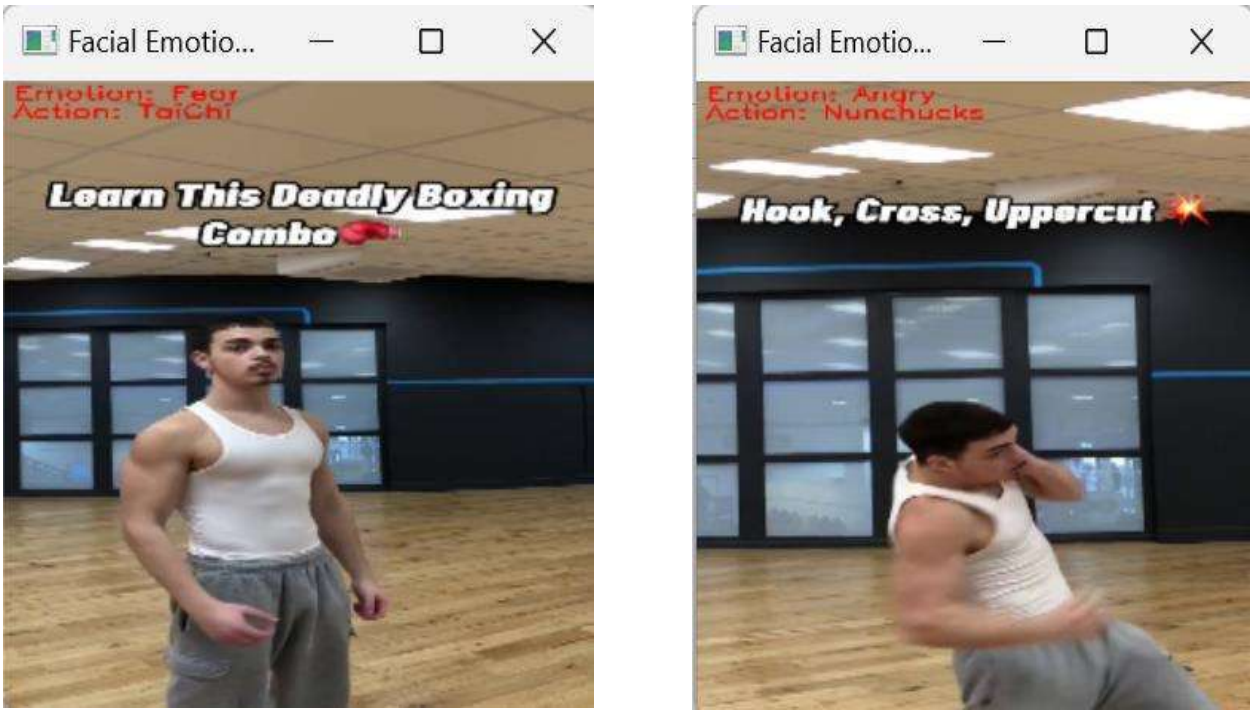

Figure. 7.3.1. Results of Combining Facial Emotion and Action Recognition in Real-Time Video Analysis

CHAPTER 8

8. CONCLUSION AND FUTURE HORIZONS

This research project has successfully developed an integrated system that combines facial emotion recognition and action recognition models to interpret human body language and facial expressions for enhanced human-machine interaction. By leveraging deep learning techniques and large-scale datasets like FER2013 and UCF101, the study has demonstrated the potential of machine learning in comprehending the richness of nonverbal communication.

The facial emotion recognition model, built using a Convolutional Neural Network (CNN) architecture and the FER2013 dataset, achieved a mean accuracy of 0.7533 (75.33%) with early stopping after 70 epochs. The training process took 472.30 seconds, showcasing the efficiency of the model in learning and predicting emotional states from facial expressions. The action recognition model, on the other hand, utilized the UCF101 dataset and a Convolutional Neural Network (CNN) architecture, achieving an impressive mean accuracy of 0.9999 (99.99%) with early stopping after 50 epochs. The training time for this model was 629.02 seconds, demonstrating its effectiveness in capturing and classifying complex actions in video streams.

By integrating these two models, the system can simultaneously recognize facial emotions and actions, providing a comprehensive understanding of human behavior in real-time. The fusion of these modalities enables a more holistic interpretation of human communication, going beyond spoken words and capturing the nuances of body language and facial expressions. This integrated approach has the potential to revolutionize various applications, such as human-computer interaction, emotion-based user interfaces, and behavior analysis in healthcare, entertainment, and security domains.

The key strengths of this research lies in its ability to handle the challenges and limitations encountered during the development process. The study addressed issues related to data preprocessing, model architecture design, and training strategies to ensure the effectiveness and robustness of the final system. The use of techniques like early stopping and cross-validation further enhanced the models' performance and generalization capabilities.

Despite the promising results, there are still opportunities for improvement and future research. One potential area for enhancement is the incorporation of additional modalities, such as speech and text, to provide a more complete understanding of human communication. By integrating these modalities with the existing facial emotion and action recognition models, the system can gain a deeper understanding of the context and intent behind human behavior.

Another direction for future research is the exploration of transfer learning and domain adaptation techniques. These approaches can help leverage pre-trained models and adapt them to specific domains or tasks, reducing the need for large-scale datasets and training from scratch. This can lead to more efficient and cost-effective development of human behavior analysis systems.

Furthermore, the deployment of the integrated system in real-world scenarios can provide valuable insights into its practical applications and limitations. Collaborations with industry partners and end-users can help identify specific use cases and tailor the system to meet their needs, ensuring its relevance and impact in various domains.

Concluding, this research project hustles progress advancing the human-engine relations by creating an unified system discerning human body language and face words. The mix of face emotion grasping and motion noticing patterns unveil the might of machine learning in grasping the intricacies of nonverbal conversation. As next steps unfold, sustained probing and creativity in this domain could drive towards smarter and more caring machines grasping and engaging with humans, opening routes for a world where human devises.

REFERENCES

[1]. Mehrotra, M., Singh, K.P. and Singh, Y.B., 2024, March. Facial Emotion Recognition and Detection Using Convolutional Neural Networks with Low Computation Cost. In *2024 2nd International Conference on Disruptive Technologies (ICDT)* (pp. 1349-1354). IEEE.

[2]. Jiao, Z., 2022, April. Research on multimodal human-computer interaction technology based on audiovisual fusion. In *2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP)* (pp. 1378-1381). IEEE.

[3]. Yu, J., Chen, K. and Xia, R., 2022. Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis. *IEEE Transactions on Affective Computing*.

[4]. Tag, B., 2023, March. Keynote: Hooked on a Feeling-Challenges and Opportunities of Emotion Research in Human-Computer Interaction. In *2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)* (pp. 256-256). IEEE.

[5]. Chen, R., Zhou, W., Li, Y. and Zhou, H., 2022. Video-based cross-modal auxiliary network for multimodal sentiment analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(12), pp.8703-8716.

[6]. Pandya, D., Pathak, R., Kumar, V., Jain, A., Jain, A. and Mursleen, M., 2023, May. Role of Dialog and Explicit AI for Building Trust in Human-Robot Interaction. In *2023 International Conference on Disruptive Technologies (ICDT)* (pp. 745-749). IEEE.

[7]. El Boudouri, Y. and Bohi, A., 2023, September. EmoNeXt: an Adapted ConvNeXt for Facial Emotion Recognition. In *2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP)* (pp. 1-6). IEEE.

[8]. Cai, Z., Gao, H., Li, J. and Wang, X., 2022, February. Deep learning approaches on multimodal sentiment analysis. In *2022 IEEE International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA)* (pp. 1127-1131). IEEE.

[9]. Fard, A.P. and Mahoor, M.H., 2022. Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild. *IEEE Access*, *10*, pp.26756-26768.

[10]. Lommatzsch, A., Llanque, B., Rosenberg, V.S., Tahir, S.A.M., Boyadzhiev, H.D. and Walny, M., Combining Information Retrieval and Large Language Models for a Chatbot that Generates Reliable, Natural-style Answers.

[11]. Ma, S., Bargal, S.A., Zhang, J., Sigal, L. and Sclaroff, S., 2017. Do less and achieve more: Training cnns for action recognition utilizing action images from the web. *Pattern Recognition*, *68*, pp.334-345.

[12]. Xie, Y., 2024. Deep Learning Approaches for Human Action Recognition in Video Data. *arXiv preprint arXiv:2403.06810*.

[13]. Shi, C. and Liu, S., 2024. Human action Recognition with Transformer based on Convolutional Features.

[14]. Yang, H., Yuan, C., Xing, J. and Hu, W., 2017, September. SCNN: Sequential convolutional neural network for human action recognition in videos. In *2017 IEEE International Conference on Image Processing (ICIP)* (pp. 355-359). IEEE.

[15]. Yang, H., Zhang, J., Li, S., Lei, J. and Chen, S., 2018. Attend it again: Recurrent attention convolutional neural network for action recognition. *Applied Sciences*, *8*(3), p.383.

[16]. Zhao, H., Xue, W., Li, X., Gu, Z., Niu, L. and Zhang, L., 2020. Multi-mode neural network for human action recognition. *IET Computer Vision*, *14*(8), pp.587-596.

[17]. Wang, L., Xu, Y., Cheng, J., Xia, H., Yin, J. and Wu, J., 2018. Human action recognition by learning spatio-temporal features with deep neural networks. *IEEE access*, *6*, pp.17913-17922.

[18]. Kim, J.H. and Won, C.S., 2020. Action recognition in videos using pre-trained 2D convolutional neural networks. *IEEE Access*, *8*, pp.60179-60188.

[19]. Jiaxin, Y., Fang, W. and Jieru, Y., 2021, March. A review of action recognition based on convolutional neural network. In *Journal of Physics: Conference Series* (Vol. 1827, No. 1, p. 012138). IOP Publishing.

[20]. Memon, F.A., Memon, M.H., Halepoto, I.A., Memon, R. and Bhangwar, A.R., 2024. Action Recognition in videos using VGG19 pre-trained based CNN-RNN Deep Learning Model. *VFAST Transactions on Software Engineering*, *12*(1), pp.46-57.

[21]. Pham, L., Vu, T.H. and Tran, T.A., 2021, January. Facial expression recognition using residual masking network. In *2020 25Th international conference on pattern recognition (ICPR)* (pp. 4513-4519). IEEE.

[22]. Shan, C., Gong, S. and McOwan, P.W., 2005, September. Robust facial expression recognition using local binary patterns. In *IEEE International Conference on Image Processing 2005* (Vol. 2, pp. II-370). IEEE.

[23]. Pecoraro, R., Basile, V. and Bono, V., 2022. Local multi-head channel self-attention for facial expression recognition. *Information*, *13*(9), p.419.

[24]. Khaireddin, Y. and Chen, Z., 2021. Facial emotion recognition: State of the art performance on FER2013. *arXiv preprint arXiv:2105.03588*.

[25]. Vulpe-Grigoraşi, A. and Grigore, O., 2021, March. Convolutional neural network hyperparameters optimization for facial emotion recognition. In *2021 12th International Symposium on Advanced Topics in Electrical Engineering (ATEE)* (pp. 1-5). IEEE.

[26]. Minaee, S., Minaei, M. and Abdolrashidi, A., 2021. Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors*, *21*(9), p.3046.

[27]. Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.H. and Zhou, Y., 2013. Challenges in representation learning: A report on three machine learning contests. In *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20* (pp. 117-124). Springer berlin heidelberg.

[28]. Chaudhari, A., Bhatt, C., Krishna, A. and Mazzeo, P.L., 2022. ViTFER: facial emotion recognition with vision transformers. *Applied System Innovation*, *5*(4), p.80.

[29]. Sharma, K. and Chanel, G., 2023, October. Annotations from speech and heart rate: impact on multimodal emotion recognition. In *Proceedings of the 25th International Conference on Multimodal Interaction* (pp. 51-59).

[30]. Fu, J., Tan, J., Yin, W., Pashami, S. and Björkman, M., 2023, October. Component attention network for multimodal dance improvisation recognition. In *Proceedings of the 25th International Conference on Multimodal Interaction* (pp. 114-118).

[31]. Vail, A.K., Girard, J.M., Bylsma, L.M., Fournier, J., Swartz, H.A., Cohn, J.F. and Morency, L.P., 2023, October. Representation Learning for Interpersonal and Multimodal Behavior Dynamics: A Multiview Extension of Latent Change Score Models. In *Proceedings of the 25th International Conference on Multimodal Interaction* (pp. 517-526).

[32]. Zhang, S., Pan, Y. and Wang, J.Z., 2023. Learning emotion representations from verbal and nonverbal communication. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 18993-19004).

[33]. Tag, B., 2023, March. Keynote: Hooked on a Feeling-Challenges and Opportunities of Emotion Research in Human-Computer Interaction. In *2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)* (pp. 256-256). IEEE.

[34]. Pandya, D., Pathak, R., Kumar, V., Jain, A., Jain, A. and Mursleen, M., 2023, May. Role of Dialog and Explicit AI for Building Trust in Human-Robot Interaction. In *2023 International Conference on Disruptive Technologies (ICDT)* (pp. 745-749). IEEE.

[35]. Ueno, T., Sawa, Y., Kim, Y., Urakami, J., Oura, H. and Seaborn, K., 2022, April. Trust in human-ai interaction: Scoping out models, measures, and methods. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts* (pp. 1-7).

[36]. Rawal, N. and Stock-Homburg, R.M., 2022. Facial emotion expressions in human–robot interaction: A survey. *International Journal of Social Robotics*, *14*(7), pp.1583-1604.

[37]. Al-Malla, M.A., Jafar, A. and Ghneim, N., 2022. Image captioning model using attention and object features to mimic human image understanding. *Journal of Big Data*, *9*(1), p.20.

[38]. Al-Malla, M.A., Jafar, A. and Ghneim, N., 2022. Image captioning model using attention and object features to mimic human image understanding. *Journal of Big Data*, *9*(1), p.20.

[39]. Savchenko, A.V., 2022. Video-based frame-level facial analysis of affective behavior on mobile devices using EfficientNets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2359-2366).

[40]. Hwooi, S.K.W., Othmani, A. and Sabri, A.Q.M., 2022. Deep learning-based approach for continuous affect prediction from facial expression images in valence-arousal space. *IEEE Access*, *10*, pp.96053-96065.

[41]. Yu, J., Chen, K. and Xia, R., 2022. Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis. *IEEE Transactions on Affective Computing*.

[42]. Chen, R., Zhou, W., Li, Y. and Zhou, H., 2022. Video-based cross-modal auxiliary network for multimodal sentiment analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(12), pp.8703-8716.

[43]. Cai, Z., Gao, H., Li, J. and Wang, X., 2022, February. Deep learning approaches on multimodal sentiment analysis. In *2022 IEEE International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA)* (pp. 1127-1131). IEEE.

[44]. Wang, Q., Saha, K., Gregori, E., Joyner, D. and Goel, A., 2021, May. Towards mutual theory of mind in human-ai interaction: How language reflects what students perceive about a virtual teaching assistant. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1-14).

[45]. Lai, H., Wu, K. and Li, L., 2021. Multimodal emotion recognition with hierarchical memory networks. *Intelligent Data Analysis*, *25*(4), pp.1031-1045.

[46]. Wang, L., Wang, S., Qi, J. and Suzuki, K., 2021. A multi-task mean teacher for semi-supervised facial affective behavior analysis. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3603-3608).

[47]. Val-Calvo, M., Álvarez-Sánchez, J.R., Ferrández-Vicente, J.M. and Fernández, E., 2020. Affective robot story-telling human-robot interaction: exploratory real-time emotion estimation analysis using facial expressions and physiological signals. *IEEE Access*, *8*, pp.134051-134066.

[48]. Zhao, X., Huang, J., Zheng, J., Ma, Y. and Tang, H., 2020, November. A multimodal-signals-based gesture recognition method for human machine interaction. In *2020 3rd International Conference on Unmanned Systems (ICUS)* (pp. 494-499). IEEE.

[49]. Kim, T. and Lee, B., 2020, June. Multi-attention multimodal sentiment analysis. In *Proceedings of the 2020 international conference on multimedia retrieval* (pp. 436-441).

[50]. Deng, D., Chen, Z. and Shi, B.E., 2020, November. Multitask emotion recognition with incomplete labels. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)* (pp. 592-599). IEEE.

[51]. Deng, J., Pang, G., Zhang, Z., Pang, Z., Yang, H. and Yang, G., 2019. cGAN based facial expression recognition for human-robot interaction. *IEEE Access*, *7*, pp.9848-9859.

[52]. Kollias, D. and Zafeiriou, S., 2019. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*.

[53]. Jaimes, A. and Sebe, N., 2007. Multimodal human–computer interaction: A survey. *Computer vision and image understanding*, *108*(1-2), pp.116-134.

Enabling smart machines to interpret human body language and facial expression

ORIGINALITY REPORT

14%	9%	10%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	open-innovation-projects.org Internet Source	1%
2	Yassine El Boudouri, Amine Bohi. "EmoNeXt: an Adapted ConvNeXt for Facial Emotion Recognition", 2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP), 2023 Publication	1%
3	arxiv.org Internet Source	<1%
4	upcommons.upc.edu Internet Source	<1%
5	"Advances in Data-Driven Computing and Intelligent Systems", Springer Science and Business Media LLC, 2024 Publication	<1%
6	www.researchgate.net Internet Source	<1%
7	fastercapital.com	

	Internet Source	<1%
8	www2.mdpi.com Internet Source	<1%
9	dokumen.pub Internet Source	<1%
10	Rongfei Chen, Wenju Zhou, Yang Li, Huiyu Zhou. "Video-based Cross-modal Auxiliary Network for Multimodal Sentiment Analysis", IEEE Transactions on Circuits and Systems for Video Technology, 2022 Publication	<1%
11	hdl.handle.net Internet Source	<1%
12	link.springer.com Internet Source	<1%
13	kylo.tv Internet Source	<1%
14	Jun-Hwa Kim, Chee Sun Won. "Action Recognition in Videos Using Pre-Trained 2D Convolutional Neural Networks", IEEE Access, 2020 Publication	<1%
15	Submitted to Liverpool John Moores University Student Paper	<1%

16	iq.opengenus.org Internet Source	<1%
17	ebin.pub Internet Source	<1%
18	"Cryptology and Network Security with Machine Learning", Springer Science and Business Media LLC, 2024 Publication	<1%
19	discovery.researcher.life Internet Source	<1%
20	Ali Pourramezan Fard, Mohammad H. Mahoor. "Ad-Corre: Adaptive Correlation-Based Loss for Facial Expression Recognition in the Wild", IEEE Access, 2022 Publication	<1%
21	Mehdi Selem, Farah Jemili, Ouajdi Korbaa. "Deep Learning for Intrusion Detection in IoT Networks", Research Square Platform LLC, 2024 Publication	<1%
22	openaccess.tau.edu.tr Internet Source	<1%
23	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	<1%
24	deepai.org Internet Source	

		<1 %
25	mdpi-res.com Internet Source	<1 %
26	www.kluniversity.in Internet Source	<1 %
27	Submitted to University of Westminster Student Paper	<1 %
28	assets.researchsquare.com Internet Source	<1 %
29	Submitted to Birkbeck College Student Paper	<1 %
30	www.mdpi.com Internet Source	<1 %
31	Submitted to Tilburg University Student Paper	<1 %
32	pure.ulster.ac.uk Internet Source	<1 %
33	Submitted to Berlin School of Business and Innovation Student Paper	<1 %
34	Jianfei Yu, Kai Chen, Rui Xia. "Hierarchical Interactive Multimodal Transformer for Aspect-Based Multimodal Sentiment	<1 %

Analysis", IEEE Transactions on Affective Computing, 2022
Publication

35	theses.liacs.nl Internet Source	<1 %
36	vivekparasharr.github.io Internet Source	<1 %
37	"Artificial Intelligence Applications and Innovations", Springer Science and Business Media LLC, 2018 Publication	<1 %
38	"Computer Vision – ACCV 2020", Springer Science and Business Media LLC, 2021 Publication	<1 %
39	Hao Yang, Chunfeng Yuan, Junliang Xing, Weiming Hu. "SCNN: Sequential convolutional neural network for human action recognition in videos", 2017 IEEE International Conference on Image Processing (ICIP), 2017 Publication	<1 %
40	Submitted to University of Technology, Sydney Student Paper	<1 %
41	www.electricaltechnology.xyz Internet Source	<1 %
42	Aayushi Chaudhari, Chintan Bhatt, Achyut Krishna, Pier Luigi Mazzeo. "ViTFER: Facial	<1 %

	Emotion Recognition with Vision Transformers", Applied System Innovation, 2022 Publication	
43	Submitted to Georgia State University Student Paper	<1 %
44	Submitted to University of Wales Institute, Cardiff Student Paper	<1 %
45	hal.archives-ouvertes.fr Internet Source	<1 %
46	www.geeksforgeeks.org Internet Source	<1 %
47	Submitted to National Institute of Technology, Silchar Student Paper	<1 %
48	thesai.org Internet Source	<1 %
49	dai-labor.de Internet Source	<1 %
50	espace.etsmtl.ca Internet Source	<1 %
51	libweb.kpfu.ru Internet Source	<1 %
52	www.jmis.org Internet Source	

		<1%
53	Hao Yang, Chunfeng Yuan, Bing Li, Yang Du, Junliang Xing, Weiming Hu, Stephen J. Maybank. "Asymmetric 3D Convolutional Neural Networks for Action Recognition", Pattern Recognition, 2018 Publication	<1%
54	Submitted to Heriot-Watt University Student Paper	<1%
55	Lecture Notes in Computer Science, 2014. Publication	<1%
56	Moumita Sen Sarma, Kaushik Deb, Pranab Kumar Dhar, Takeshi Koshiba. "Traditional Bangladeshi Sports Video Classification Using Deep Learning Method", Applied Sciences, 2021 Publication	<1%
57	www.ijraset.com Internet Source	<1%
58	www.medrxiv.org Internet Source	<1%
59	"MI-STA 2022 Conference Proceeding", 2022 IEEE 2nd International Maghreb Meeting of the Conference on Sciences and Techniques	<1%

of Automatic Control and Computer Engineering (MI-STA), 2022
Publication

60	Jamin Rahman Jim, Md Apon Riaz Talukder, Partha Malakar, Md Mohsin Kabir, Kamruddin Nur, M.F. Mridha. "Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review", Natural Language Processing Journal, 2024 Publication	<1 %
61	Lei Wang, Yangyang Xu, Jun Cheng, Haiying Xia, Jianqin Yin, Jiaji Wu. "Human Action Recognition by Learning Spatio-Temporal Features with Deep Neural Networks", IEEE Access, 2018 Publication	<1 %
62	Submitted to Napier University Student Paper	<1 %
63	Submitted to Nottingham Trent University Student Paper	<1 %
64	Umair Ali Khan, Qianru Xu, Yang Liu, Altti Lagstedt, Ari Alamäki, Janne Kauttonen. "Exploring contactless techniques in multimodal emotion recognition: insights into diverse applications, challenges, solutions, and prospects", Multimedia Systems, 2024 Publication	<1 %

65	Submitted to CSU, San Diego State University Student Paper	<1 %
66	Submitted to Monash University Student Paper	<1 %
67	Palanichamy Naveen. "Occlusion-aware facial expression recognition: A deep learning approach", Multimedia Tools and Applications, 2023 Publication	<1 %
68	Submitted to University of Liverpool Student Paper	<1 %
69	www.codewithc.com Internet Source	<1 %
70	Submitted to Al Akhawayn University in Ifrane Student Paper	<1 %
71	Submitted to Cardiff University Student Paper	<1 %
72	Submitted to Coventry University Student Paper	<1 %
73	Jun Li, Xianglong Liu, Mingyuan Zhang, Deqing Wang. "Spatio-temporal deformable 3D ConvNets with attention for action recognition", Pattern Recognition, 2020 Publication	<1 %

74	Submitted to Southern New Hampshire University - Continuing Education Student Paper	<1 %
75	Submitted to University of Hertfordshire Student Paper	<1 %
76	Submitted to University of Portsmouth Student Paper	<1 %
77	Submitted to University of Surrey Student Paper	<1 %
78	dias.library.tuc.gr Internet Source	<1 %
79	Submitted to Institute of International Studies Student Paper	<1 %
80	S. Muhammed, J. Upadhya, S. Poudel, M. Hasan, K. Donthula, J. Vargas, J. Ranganathan, K. Poudel. "Improved Classification of Alzheimer's Disease With Convolutional Neural Networks", 2023 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), 2023 Publication	<1 %
81	koreascience.kr Internet Source	<1 %
82	Nyle Siddiqui, Thomas Reither, Rushit Dave, Dylan Black, Tyler Bauer, Mitchell Hanson. "A Robust Framework for Deep Learning	<1 %

Approaches to Facial Emotion Recognition and Evaluation", 2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML), 2022
Publication

---

83 Pipit Utami, Rudy Hartanto, Indah Soesanti. "The EfficientNet Performance for Facial Expressions Recognition", 2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 2022 <1%
Publication

---

84 Sheng Yu, Yun Cheng, Li Xie, Zhiming Luo, Min Huang, Shaozi Li. "A novel recurrent hybrid network for feature fusion in action recognition", Journal of Visual Communication and Image Representation, 2017 <1%
Publication

---

85 Yuhui Quan, Yixin Chen, Ruotao Xu, Hui Ji. "Attention with structure regularization for action recognition", Computer Vision and Image Understanding, 2019 <1%
Publication

---

86 ceur-ws.org <1%
Internet Source

---

87 eprints.bournemouth.ac.uk <1%
Internet Source

88	www.grafiati.com Internet Source	<1%
89	D Khalandar Basha, G Sunil, A. H. A. Hussein, Mukesh S, Myasar Mundher Adnan. "Multimodal sentiment analysis using Multi-Layer Fusion Convolution Neural Network", 2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIE), 2023 Publication	<1%
90	Dhruv Saluja, Harsh Kukreja, Akash Saini, Devanshi Tegwal, Preeti Nagrath, Jude Hemanth. "Analysis and comparison of various deep learning models to implement suspicious activity recognition in CCTV surveillance", Intelligent Decision Technologies, 2023 Publication	<1%
91	Sumeet Saurav, Ravi Saini, Sanjay Singh. "An integrated attention-guided deep convolutional neural network for facial expression recognition in the wild", Multimedia Tools and Applications, 2024 Publication	<1%
92	digitalcommons.unl.edu Internet Source	<1%
93	iieta.org Internet Source	

		<1 %
94	lib.opt.ac.cn Internet Source	<1 %
95	www.brnsspubhub.org Internet Source	<1 %
96	www.ijert.org Internet Source	<1 %
97	www.shs-conferences.org Internet Source	<1 %
98	"Cognitive Computing and Cyber Physical Systems", Springer Science and Business Media LLC, 2024 Publication	<1 %
99	Asifullah Khan, Zunaira Rauf, Anabia Sohail, Abdul Rehman Khan, Hifsa Asif, Aqsa Asif, Umair Farooq. "A survey of the vision transformers and their CNN-transformer based variants", Artificial Intelligence Review, 2023 Publication	<1 %
100	Ganesh Chandrasekaran, S. Dhanasekaran, C. Moorthy, A. Arul Oli. "Multimodal sentiment analysis leveraging the strength of deep neural networks enhanced by the XGBoost classifier", Computer Methods in	<1 %

Biomechanics and Biomedical Engineering, 2024
Publication

101	Ibtissam Saadi, Douglas W. cunningham, Taleb-Ahmed Abdelmalik, Abdenour Hadid, Yassin El Hillali. "Driver's facial expression recognition: A comprehensive survey", Expert Systems with Applications, 2023 Publication	<1 %
102	Lecture Notes in Computer Science, 2015. Publication	<1 %
103	Shubhanjay Pandey, Sonakshi Handoo, Yogesh. "Facial Emotion Recognition using Deep Learning", 2022 International Mobile and Embedded Technology Conference (MECON), 2022 Publication	<1 %
104	Thomas Kopalidis, Vassilios Solachidis, Nicholas Vretos, Petros Daras. "Advances in Facial Expression Recognition: A Survey of Methods, Benchmarks, Models, and Datasets", Information, 2024 Publication	<1 %
105	Tong Liu, Jing Li, Jia Wu, Bo Du, Jun Wan, Jun Chang. "Confusable facial expression recognition with geometry-aware conditional network", Pattern Recognition, 2023 Publication	<1 %

106	Vishwam Jaimini Pandya. "Comparing Handwritten Character Recognition by AdaBoostClassifier and KNeighborsClassifier", 2016 8th International Conference on Computational Intelligence and Communication Networks (CICN), 2016 Publication	<1 %
107	Yang Liu, Zhaoyang Lu, Jing Li, Tao Yang, Chao Yao. "Deep Image-to-Video Adaptation and Fusion Networks for Action Recognition", IEEE Transactions on Image Processing, 2020 Publication	<1 %
108	Yuelin Li, Jiayi Zhao, Yutang Lu. "Strengthening Emotion Recognition Algorithms: A Defense Mechanism against FGSM White-Box Attacks", 2023 6th International Conference on Algorithms, Computing and Artificial Intelligence, 2023 Publication	<1 %
109	aisel.aisnet.org Internet Source	<1 %
110	dspace.daffodilvarsity.edu.bd:8080 Internet Source	<1 %
111	eprints.kfupm.edu.sa Internet Source	<1 %
112	export.arxiv.org Internet Source	<1 %

113	iarjset.com Internet Source	<1 %
114	inemg.com Internet Source	<1 %
115	ksascholar.dri.sa Internet Source	<1 %
116	mranta-ai.github.io Internet Source	<1 %
117	www.albion.edu Internet Source	<1 %
118	www.arxiv-vanity.com Internet Source	<1 %
119	www.biorxiv.org Internet Source	<1 %
120	www.techscience.com Internet Source	<1 %
121	www.um.edu.mt Internet Source	<1 %
122	"Computer Vision – ECCV 2016 Workshops", Springer Nature, 2016 Publication	<1 %
123	Benjamin Tag. "Keynote: Hooked on a Feeling - Challenges and Opportunities of Emotion Research in Human-Computer Interaction", 2023 IEEE International Conference on	<1 %

Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), 2023
Publication

---

124 Jason C. Hung, Kuan-Cheng Lin, Nian-Xiang Lai. "Recognizing learning emotion based on convolutional neural networks and transfer learning", Applied Soft Computing, 2019 <1 %
Publication

---

125 Jun-Hwa Kim, Chee Sun Won. "Audio-Visual Action Recognition Using Transformer Fusion Network", Applied Sciences, 2024 <1 %
Publication

---

126 Kunhong Xiong, Linbo Qing, Lindong Li, Li Guo, Yonghong Peng. "Facial expression recognition based on local–global information reasoning and spatial distribution of landmark features", The Visual Computer, 2024 <1 %
Publication

---

127 Shuvendu Roy, Ali Etemad. "Analysis of Semi-Supervised Methods for Facial Expression Recognition", 2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII), 2022 <1 %
Publication

---

128 Zisheng Cai, Han Gao, Jiaye Li, Xinyi Wang. "Deep Learning Approaches on Multimodal <1 %

Sentiment Analysis", 2022 IEEE International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA), 2022
Publication

---

129 "Ambient Intelligence – Software and Applications – 14th International Symposium on Ambient Intelligence", Springer Science and Business Media LLC, 2023 <1%
Publication

---

130 "Computer Vision – ECCV 2022", Springer Science and Business Media LLC, 2022 <1%
Publication

---

131 "Intelligent Robotics and Applications", Springer Science and Business Media LLC, 2022 <1%
Publication

---

132 Alex D. Torres, Hao Yan, Armin Haj Aboutalebi, Arun Das, Lide Duan, Paul Rad. "Patient Facial Emotion Recognition and Sentiment Analysis Using Secure Cloud With Hardware Acceleration", Elsevier BV, 2018 <1%
Publication

---

133 Alireza Ghorbanali, Mohammad Karim Sohrabi. "A comprehensive survey on deep learning-based approaches for multimodal sentiment analysis", Artificial Intelligence Review, 2023 <1%
Publication

134	Di Wang, Changning Tian, Xiao Liang, Lin Zhao, Lihuo He, Quan Wang. "Dual-Perspective Fusion Network for Aspect-based Multimodal Sentiment Analysis", IEEE Transactions on Multimedia, 2024 Publication	<1%
135	Haodong Yang, Jun Zhang, Shuohao Li, Jun Lei, Shiqi Chen. "Attend It Again: Recurrent Attention Convolutional Neural Network for Action Recognition", Applied Sciences, 2018 Publication	<1%
136	Haohua Zhao, Weichen Xue, Xiaobo Li, Zhangxuan Gu, Li Niu, Liqing Zhang. "Multi-mode neural network for human action recognition", IET Computer Vision, 2020 Publication	<1%
137	Roberto Pecoraro, Valerio Basile, Viviana Bono. "Local Multi-Head Channel Self-Attention for Facial Expression Recognition", Information, 2022 Publication	<1%
138	Sanjeda Sara Jennifer, Mahbub Hasan Shamim, Ahmed Wasif Reza, Nazmul Siddique. "Sickle cell disease classification using deep learning", Heliyon, 2023 Publication	<1%

139 Sumeet Saurav, Ravi Saini, Sanjay Singh. "A dual-channel ensembled deep convolutional neural network for facial expression recognition in the wild", Computational Intelligence, 2023 <1%

Publication

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

PHOTOGRAPH ALONG WITH GUIDE