UNIVERSITY OF PISA

MASTER'S DEGREE IN COMPUTER ENGINEERING

Industrial Applications

MoodPilot

Professors:

Pierfrancesco Foglia

Antonio Cosimo Prete

Students:

Giovanni Ligato

Giuseppe Soriano

ACADEMIC YEAR 2024/2025

Abstract

In this study, we present a system for evaluating user satisfaction with autonomous and manual driving experiences through Facial Emotion Recognition (FER). A detailed analysis was conducted to identify and benchmark state-of-the-art FER models, including DeepFace, EmoNet, HSEmotionONNX, RMN, and Vision Transformer (ViT), with additional tests performed on a Raspberry Pi 3B+ as a reference for potential car controller deployment. Alongside this analysis, we developed a frontend and backend system to facilitate data collection for future training of a predictive model. The system allows users to manually respond to a form and undergo video-based analysis of their driving experience. The ultimate goal is to train a model that can autonomously infer user responses based on collected data, paving the way for an adaptive system to evaluate driving satisfaction efficiently.

Index

1. Introduction

Emotions play a fundamental role in shaping human cognition, behavior, and decision-making. They influence how individuals perceive and interact with their surroundings, making them critical in contexts that require human-machine interaction. Recent advancements in technology have enabled the integration of emotion recognition into various applications, ranging from healthcare to social robotics.

In the automotive domain, emotions are particularly relevant for assessing user satisfaction in different driving modes, such as autonomous driving and manual control. Our system aims to provide an innovative service that evaluates passenger experiences automatically, based on their perceived emotions. This service can be utilized by companies offering ride-hailing and driving services, enabling them to assess the quality of drivers or autonomous driving systems and improve customer satisfaction. By analyzing passengers' emotional states during a ride, the system delivers insights that help identify strengths and areas for improvement, fostering better service quality and user trust.

Facial Emotion Recognition (FER) emerges as a key technology in this context. By analyzing facial expressions, FER systems infer users' emotional states in real-time, providing an objective evaluation of their experiences. This capability is especially valuable in ride-hailing services, where customer feedback is essential for evaluating and enhancing driving performance.

However, implementing FER in constrained environments like in-car systems poses unique challenges. Factors such as limited computational resources, varying environmental conditions (e.g., lighting, seating positions), and privacy concerns must be addressed to create effective solutions. This study investigates these challenges, building the foundation for a system that leverages FER to improve passenger satisfaction and trust in driving technologies.

1.1. Emotion Recognition: Applications and Challenges

Emotions are complex psychological states that encompass subjective experiences, physiological responses, and behavioral expressions. They play a crucial role in shaping human cognition, influencing decision-making, perception, and social interactions. Understanding and categorizing emotions has been a focus of psychological research, leading to the development of various theoretical models. These frameworks have laid the foundation for technologies like Facial Emotion Recognition (FER), which aim to interpret emotions from observable cues such as facial expressions.

1.1.1. Emotion Classifications

Two major models dominate the study of emotion categorization:

- **Discrete Emotion Theory**: This theory suggests that humans experience a set of universal, basic emotions. Pioneered by Paul Ekman, research has identified six fundamental emotions—anger, disgust, fear, happiness, sadness, and surprise—each linked to distinct facial expressions and physiological responses. These emotions are universally recognizable across cultures, supporting the notion of a biological basis for emotional expression [1].
- **Dimensional Models**: These models conceptualize emotions as points within a continuous space rather than discrete categories. One prominent example is the *Circumplex Model*,

proposed by James Russell, which organizes emotions along two axes:

- **Valence**: The positivity or negativity of an emotion (e.g., happiness is positive, sadness is negative).
- **Arousal**: The intensity or activation level of an emotion (e.g., excitement is high-arousal, calmness is low-arousal) [2].

These frameworks provide structured ways to interpret and classify emotions, offering significant utility for applications like FER in automotive contexts.

1.1.2. Discrete Emotion Theory: Universality and Characteristics

The Discrete Emotion Theory, rooted in Charles Darwin's work, posits that emotions have evolutionary purposes, aiding survival and adaptation. Ekman's cross-cultural studies confirmed the universality of six basic emotions, which are recognized and expressed consistently across populations.

Each of these basic emotions has unique characteristics:

- **Anger**: Associated with perceived threats or injustices, marked by physiological arousal and narrowed focus.
- **Disgust**: Often a response to harmful or offensive stimuli, with distinct facial cues like nose wrinkling.
- **Fear**: Triggers the fight-or-flight response, enhancing focus and readiness for action.
- **Happiness**: Signifies satisfaction and well-being, expressed through smiles and other positive indicators.
- **Sadness**: Linked to loss or disappointment, characterized by lowered energy.
- **Surprise**: Prompted by unexpected events, facilitating rapid attention shifts with widened eyes and raised eyebrows.

Ekman's studies demonstrated that these emotions are universally recognized through facial expressions, even in cultures isolated from global influences. For instance, widened eyes and raised eyebrows universally signal surprise, while smiles denote happiness. This universality suggests a biological basis for basic emotions, enabling effective nonverbal communication.

1.1.3. The Circumplex Model: Valence and Arousal

Developed by James Russell, the Circumplex Model (depicted in Figure 1) maps emotions based on two dimensions:

- **Valence**: Represented on the horizontal (X) axis, valence measures the positivity or negativity of an emotion. Positive valence indicates pleasant emotions (e.g., happiness), while negative valence corresponds to unpleasant emotions (e.g., sadness).

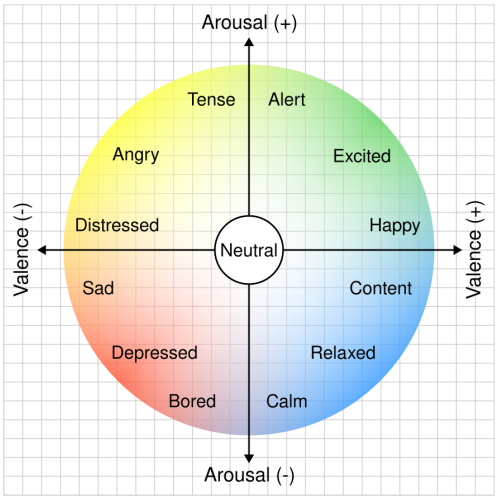


Figure 1: The Circumplex Model of Emotion maps emotions within a two-dimensional space defined by valence (X-axis) and arousal (Y-axis). Positive and negative valence correspond to pleasant and unpleasant emotions, respectively, while arousal indicates emotional intensity. Image taken from en.wikipedia.org/wiki/Emotion_classification

- **Arousal**: Represented on the vertical (Y) axis, arousal gauges the intensity or activation level of an emotion, ranging from low (calmness) to high (excitement).

By plotting emotions within this two-dimensional space, the Circumplex Model illustrates how different emotions relate to one another. For instance, emotions like excitement (high arousal, positive valence) and calmness (low arousal, positive valence) are positioned accordingly, highlighting the continuous nature of emotional experiences.

These models are vital for fields like Facial Expression Recognition (FER), as they offer structured frameworks for interpreting and categorizing human emotions based on observable cues, thereby advancing human-computer interaction and social robotics.

1.1.4. Challenges in Emotion Recognition

While these models provide robust theoretical foundations, implementing FER in real-world automotive settings introduces several challenges:

- **Environmental Variability**: In-car lighting and passenger positioning can degrade facial analysis quality.

- **Computational Constraints**: Many FER systems must operate on low-power devices like Raspberry Pi 3, balancing accuracy and efficiency.

- **Privacy Concerns**: FER systems must address ethical issues surrounding the collection and processing of sensitive facial data.

By leveraging these emotion classification models, this study explores the feasibility of FER for evaluating passenger satisfaction and improving user trust in ride-hailing and autonomous driving technologies.

1.2. Scope of the Study

This study aims to explore the potential of Facial Emotion Recognition (FER) in evaluating passenger experiences during driving scenarios, with the ultimate goal of improving user satisfaction and trust in ride-hailing and autonomous driving services. To achieve this, the project focuses on the following key objectives:

- **Benchmarking State-of-the-Art FER Models** A comprehensive analysis of FER models is conducted to evaluate their performance and suitability for deployment in automotive contexts. The models under investigation include DeepFace, EmoNet, HSEmotionONNX, RMN, and Vision Transformer (ViT). Each model is assessed for its accuracy, robustness, and efficiency, particularly in constrained environments.
- **Adaptation for Edge Devices** The feasibility of deploying FER models on low-power devices, such as the Raspberry Pi 3, is explored. This device serves as a reference for potential car controllers, where computational resources are limited, and real-time processing is essential.
- **Development of a Data Collection System** A dual-component system, comprising frontend and backend modules, was developed to facilitate the collection of data from passengers. Users are asked to complete a form about their driving experience, while facial expressions are recorded during the ride. This dataset serves as a foundation for training predictive models.
- **Future Model Training** The collected data will be used to train a model capable of autonomously responding to form questions based on passengers' inferred emotions. This development aims to bridge the gap between user feedback and adaptive driving systems, enabling real-time evaluation and adjustments.
- **Addressing Key Challenges** The study addresses challenges such as computational constraints and privacy concerns. These considerations ensure that the developed system is not only effective but also applicable in real-world scenarios.

By integrating these objectives, the study provides a solid foundation for emotion-aware automotive systems.

2. State of the Art

This section provides a comprehensive review of the **State of the Art**, highlighting advancements and challenges in key technological domains. It examines the limitations of two-way rating systems within ride-hailing platforms, explores the integration of Facial Emotion Recognition (FER) technologies to enhance user experience and safety in automotive contexts, and evaluates the deployment of FER systems on edge devices, emphasizing their potential for real-time, privacy-preserving applications in resource-constrained environments.

2.1. Two-ways Rating System Limitations in Ride-Hailing Services

Reputation systems play a pivotal role in ride-sharing platforms, serving as a self-regulation mechanism that aims to ensure service quality and accountability. The two-way rating system, where both drivers and passengers evaluate each other, has become the industry standard; however, its **limitations** have been widely discussed in the literature. Studies highlight that while such systems are designed to exclude low-rated users, their effectiveness is often constrained by the platform's structural and incentive-based shortcomings. For instance, compliance with platform rules remains imperfect even under ideal rating conditions, and the perceived cost of rebuilding a damaged reputation may not always be sufficient to incentivize quality service delivery. Furthermore, the *opacity* of reputation algorithms, while controversial, can paradoxically enhance compliance by preserving an element of uncertainty about how penalties are applied [3].

Despite the potential of ride-sharing systems to promote sustainability, reduce car usage, and increase vehicle occupancy, they face significant barriers that hinder their development and adoption. A systematic review identifies **economic**, **technological**, **behavioral**, and **regulatory** challenges as critical factors affecting user participation and system efficiency. User-specific attributes, such as sociodemographics and location, also play a role in shaping ride-sharing behavior, although their influence shows *mixed results* across different studies. Additionally, the lack of integration with public transport and concerns over safety further limit scalability [4]. At the core of these challenges lies the reliance on rating systems, which often fail to capture the *complexity* of service quality evaluation. Ratings may be biased due to **fear of retaliation** or unclear criteria, reducing their fairness and accuracy as a quality metric. Moreover, differences in operational contexts, payment methods, and user trip purposes complicate the creation of standardized reputation mechanisms, as seen in cross-national analyses [5].

To address these limitations, researchers have proposed integrating reputation systems with real-time contextual data, such as trip purpose and environmental factors, to provide a more nuanced evaluation of service quality. The concept of cross-platform **reputation portability** has also emerged as a potential solution, allowing users to transfer their reputation across different platforms, thereby incentivizing better service standards. Furthermore, targeting specific user types, such as **solo work commuters** or recreational users, could enhance the effectiveness of ride-sharing services by tailoring offerings to diverse needs [3] [5]. While the two-way rating system remains a foundational element of ride-sharing platforms, these findings underscore the need for more **adaptive** and **transparent** mechanisms that can overcome current inefficiencies and foster greater trust and participation among users.

2.2. Facial Emotion Recognition (FER) in Automotive Contexts

The ability to recognize and interpret human emotions has emerged as a critical component in the development of emotion-aware automotive systems. Facial Emotion Recognition (FER) offers significant potential in this area, enabling vehicles to adapt to users' emotional states and enhance the overall driving experience. Recent research underscores the importance of FER in both traditional and autonomous vehicles, addressing various applications and challenges in human-vehicle interaction.

Driving often elicits **emotional states** such as stress, frustration, or anger, which can influence safety, comfort, and decision-making on the road. A comprehensive survey of emotion recognition in automotive contexts highlights a preference for monitoring high-arousal, negative-valence states, which are strongly associated with driver behavior and road safety risks. Common methodologies include the use of multimodal signals such as cardiac and electrodermal activity, speech analysis, and facial expressions. These are processed using supervised machine learning models to infer underlying emotional states [6].

As the focus shifts from drivers to passengers in the context of autonomous vehicles (AVs), FER applications are expanding to include enhancing passenger experience. Emotion-aware systems can adapt driving styles to create more pleasant journeys, tailor infotainment systems to individual preferences, and manage group affect to ensure collective comfort. Multimodal approaches combining visual and audio inputs are essential in these scenarios, addressing challenges such as **context awareness**, multi-passenger diarization, and personalization. Researchers also emphasize the importance of **explainability** and privacy in developing these systems [7].

In connected and automated vehicles (CAVs), affective interaction frameworks are becoming increasingly sophisticated. These frameworks integrate knowledge from various disciplines, including human-machine interaction, automotive engineering, and communication. FER plays a pivotal role in facilitating **human-vehicle-road systems** by detecting, regulating, and responding to user emotions in real-time. Such systems aim to improve the acceptance, safety, and comfort of CAVs, offering a more natural and enjoyable user experience. However, the design of these frameworks demands a deep understanding of multimodal emotional expressions and their interplay with vehicular networking. Promising research directions include advancing multimodal emotion detection, refining emotion regulation strategies, and exploring applications for group interactions and dynamic scenarios [8].

Despite these advances, several challenges remain. Accurate FER systems must contend with variability in lighting, seating positions, and cultural differences in emotional expression. Moreover, ensuring user **privacy** and implementing ethical data handling practices are critical to fostering trust in emotion-aware automotive systems. Future research should focus on **holistic modelling** approaches that integrate multiple data streams while addressing the unique constraints of vehicular environments.

By incorporating FER into automotive systems, the industry moves closer to achieving adaptive, user-centered vehicles that enhance safety, comfort, and overall satisfaction. The integration of emotion recognition technologies into intelligent vehicles not only addresses immediate concerns like road safety but also paves the way for innovative user experiences in next-generation mobility solutions.

2.3. Facial Emotion Recognition (FER) on Edge Devices

In "Using emotion recognition and temporary mobile social network in on-board services for car passengers" [9], the authors evaluate the deployment of facial emotion recognition (FER) systems on Raspberry Pi 4 B devices, showcasing the potential for real-time emotion detection in edge-based applications. The study focuses on leveraging edge computing to address the limitations of cloud-based solutions, such as privacy concerns, reliance on connectivity, and the complexity of maintaining connected infrastructures. The proposed system architecture includes a face detection (FD) module to locate and extract faces from input images, followed by a FER module to classify emotions.

Several FD algorithms were initially explored using frameworks like OpenCV and Darknet. OpenCV's Haar Cascade and Improved Local Binary Patterns (ILBP) were evaluated for their execution time, but ultimately Yoloface-500k v2, a lightweight model based on YOLOv3, was chosen for its superior balance of accuracy and performance. For the FER module, two models were tested: DeepFace, a lightweight Python library capable of running directly on the Raspberry Pi, and Emonet, which required the use of a Neural Compute Stick 2 (NCS2) accelerator due to its higher computational demand.

Two versions of the system were developed. The first integrated both the FD and DeepFace FER models on the Raspberry Pi, providing efficient performance suitable for real-time applications. The second utilized Yoloface for FD on the Raspberry Pi and offloaded the FER task to the NCS2 accelerator for running Emonet. Since Emonet took over 10 seconds per frame on the Raspberry Pi alone, the use of the NCS2 significantly improved processing time. Additional optimizations, such as asynchronous pipelining, further reduced latency by overlapping face detection and emotion recognition tasks for successive frames.

The study concludes that systems leveraging lightweight FD models like Yoloface-500k v2 and supported by accelerators such as the NCS2 can enable real-time FER even on resource-constrained edge devices. This approach highlights the feasibility of deploying FER for privacy-sensitive applications, including automotive environments, where the system can adapt services dynamically based on detected user moods.

3. System Description

The system architecture for the MoodPilot project is designed to seamlessly integrate data collection, processing, and user interaction across its components. The following describes the main components of the system:

Database: The system utilizes MongoDB as the primary database to store structured and unstructured data. This includes user data, ride logs, emotional feedback, and FER (Facial Emotion Recognition) model outputs. MongoDB's flexibility and scalability ensure efficient handling of data streams and support for real-time updates.

Backend: The backend is implemented using Flask, a lightweight and extensible Python web framework. It acts as the central hub, managing requests from the frontend and car controllers. Flask handles:

- Communication with the database for data storage and retrieval.
- Processing and aggregation of data to provide meaningful insights.
- Integration with the model used to respond automatically to forms based on emotional data, reducing the computational load on the car controller.
- Secure authentication and session management for mobile app users.

Car Controller: The car controller is an embedded system integrated within the vehicle, responsible for running the FER model. It processes live video streams or images captured within the vehicle to evaluate the emotional states of passengers. The FER model provides real-time feedback on passenger satisfaction, which is logged in the database and analyzed for service improvement. The car controller communicates with the backend to sync evaluation data. By offloading the execution of the form-response model to the backend, the car controller focuses solely on the FER model, ensuring optimal performance.

Frontend: The frontend is a mobile application designed to provide a user-friendly interface for passengers. The app includes features such as:

- Booking ride-hailing services.
- Viewing and tracking ongoing rides.
- Providing feedback about the ride experience.
- Accessing a history of previous rides and satisfaction scores.
- Visualizing real-time metrics of ride quality, such as safety and comfort scores derived from FER and sensor data. The mobile app communicates with the backend through secure APIs to ensure seamless service.

Security and Privacy: The system incorporates robust security measures to protect user data and maintain privacy. This includes:

- End-to-end encryption for data transmission.
- Anonymization of sensitive passenger data during FER processing.
- Regular audits and compliance with data protection regulations.

Overall, the system is built with scalability and modularity in mind, allowing for future extensions, such as integrating additional sensors, refining FER models for enhanced performance, or supporting advanced analytics and reporting tools.

4. Facial Expression Recognition (FER) Performance Analysis

Facial Expression Recognition (FER) systems are critical in understanding human emotions through visual cues, offering applications in diverse fields such as human-computer interaction, mental health assessment, and automotive contexts. This section delves into the datasets and models used for training and evaluating FER systems, with a particular emphasis on their performance and suitability for real-world applications.

4.1. Datasets

The datasets described in this section are central to the development and benchmarking of FER models. They are referenced throughout the document to indicate their specific use cases in training and evaluation.

4.1.1. FER2013

FER2013 [10] is a widely-used dataset for facial expression recognition, containing 48x48 pixel grayscale images of faces. Each image has been automatically aligned and scaled so that the face is centered and occupies a similar amount of space across samples. The dataset includes seven emotion categories:

- **Angry (0)**
- **Disgust (1)**
- **Fear (2)**
- **Happy (3)**
- **Sad (4)**
- **Surprise (5)**
- **Neutral (6)**

FER2013 comprises 28,709 training samples and 3,589 test samples. The task involves classifying each image into one of the seven predefined emotional categories. The dataset's relatively small resolution and balanced scaling make it a suitable benchmark for evaluating FER models, although its limited variability in facial contexts poses challenges for generalization.

4.1.2. AffectNet

AffectNet [11] addresses the scarcity of large-scale, annotated facial expression datasets, especially those covering both discrete emotion categories and continuous affect dimensions (valence and arousal). AffectNet is currently the largest facial expression dataset, containing over 1 million facial images collected from the internet using 1,250 emotion-related keywords in six languages. Of these, approximately 440,000 images were manually annotated.

- **Categorical Model**: The dataset includes eight discrete emotion labels:
  - **Angry**
  - **Disgust**
  - **Fear**
  - **Happy**
  - **Sad**

  - **Surprise**
  - **Neutral**
  - **Contempt**
- **Dimensional Model**: Each image is also annotated for valence (positive to negative emotional intensity) and arousal (level of emotional activation).

AffectNet's diverse and comprehensive annotations make it a versatile resource for FER research in both categorical and dimensional frameworks. The dataset's scale and annotation quality support the development of robust models capable of handling real-world variability in facial expressions.

4.2. Models

In the following the primary models used in FER research and their performance on the aforementioned datasets are presented. Each model is described briefly, including its architecture, dataset usage, and key performance metrics.

4.2.1. DeepFace

DeepFace is a lightweight and versatile Python framework designed for face recognition and facial attribute analysis, including emotion detection, age, gender, and ethnicity prediction [12]. It serves as a hybrid platform that wraps state-of-the-art models such as VGG-Face, FaceNet, OpenFace, DeepFace, DeepID, ArcFace, and others, enabling a wide range of facial analysis applications.

The framework implements a modern facial recognition pipeline, automating key steps such as face detection, alignment, normalization, representation, and verification. This design ensures that users can perform complex facial analysis tasks with minimal configuration, leveraging DeepFace's intuitive API. For emotion recognition specifically, DeepFace integrates robust pre-trained models capable of classifying facial expressions into categories such as anger, happiness, sadness, and more.

DeepFace's modular architecture supports various backends for face detection, including OpenCV, MTCNN, and RetinaFace, providing flexibility in balancing speed and accuracy. For emotion recognition, it uses facial embeddings generated by convolutional neural networks (CNNs), allowing efficient and accurate classification. The framework also supports multiple distance metrics, such as cosine similarity and Euclidean distance, to compute facial similarities in verification tasks.

By offering high-level abstractions and combining cutting-edge techniques, DeepFace achieves competitive performance on datasets like **FER2013**, with an accuracy of **57.42%**. Its ease of integration and ability to leverage multiple advanced models make it an essential tool for both research and real-world applications in facial expression recognition and analysis.

4.2.2. HSEmotionONNX

HSEmotionONNX is a collection of ONNX-compatible models designed for efficient and accurate facial emotion recognition, developed by Andrey Savchenko during his research at HSE University and Sber AI Lab [13]. These models were initially pre-trained on the VGGFace2 dataset for face identification and later fine-tuned on AffectNet, achieving state-of-the-art performance in recognizing both categorical emotions and valence-arousal dimensions.

The library supports multiple lightweight models, including variations of EfficientNet, which balance accuracy and computational efficiency. Notably, the **enet_b0_8_best_vgaf** model demonstrates robust performance with accuracies of 61.32% for 8 emotion classes and 64.57% for 7 classes, with an average inference time of 59 ŏ0b1 26 ms and a model size of 16 MB. Similarly, the **enet_b2_8** model achieves higher accuracy (63.03% for 8 classes and 66.29% for 7 classes) but at the cost of increased computational demand, requiring 191 ŏ0b1 18 ms per inference and occupying 30 MB of memory.

These models excel in scenarios requiring real-time emotion recognition on resource-constrained devices, such as mobile platforms and edge computing environments. Their efficiency is further enhanced by leveraging ONNX's cross-platform compatibility, allowing deployment across various frameworks and devices. The preferred model, **enet_b0_8_best_vgaf**, offers an optimal trade-off between speed and accuracy, making it suitable for applications where high frame rates and responsiveness are critical.

HSEmotionONNX also provides tools for analyzing emotions at the batch level, supporting applications like video-level emotion tracking and multi-user affective computing. The library's flexibility, combined with its high performance, makes it a powerful tool for advancing research and development in emotion recognition technologies.

4.2.3. Vision Transformer (ViT) for Facial Expression Recognition

The Vision Transformer (ViT) for Facial Expression Recognition is a cutting-edge model fine-tuned on the FER2013 dataset for the task of emotion recognition [14]. It leverages the **vit-base-patch16-224-in21k** architecture, a pre-trained Vision Transformer initially trained on ImageNet and adapted for facial emotion classification. The model achieves notable accuracies of **71.13% on the validation set** and **71.16% on the test set**, demonstrating its effectiveness in recognizing seven discrete emotion categories: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral.

The ViT model processes input images by splitting them into fixed-sized patches. These patches undergo an embedding phase initiated by a convolutional layer with a 16x16 kernel and a stride of 16x16. The resulting embeddings are then enriched with positional embeddings and projected into a 768-dimensional feature space. This sequence of embedded patches is processed through 11 Transformer Encoder layers, allowing the model to capture complex spatial relationships within the images. For the emotion classification task, the final output layer is a fully connected linear layer with eight dimensions, corresponding to the emotional categories.

To prepare the images for input into the model, several preprocessing steps are applied:

- **Resizing:** Images are resized to match the input dimensions expected by the model.
- **Normalization:** Pixel values are scaled to a specific range to ensure consistency.
- **Data Augmentation:** Random transformations, such as rotations, flips, and zooms, are applied during training to increase dataset variability and improve model generalization.

While the ViT model demonstrates strong performance, its accuracy is subject to the quality and diversity of the FER2013 dataset. The dataset's inherent biases and limited variability may influence the model's ability to generalize effectively to unseen data. Future enhancements could include training on more diverse datasets or incorporating additional data augmentation techniques.

The Vision Transformer architecture's ability to leverage global attention mechanisms within images

positions it as a promising approach for emotion recognition, combining state-of-the-art performance with a scalable design.

4.2.4. Residual Masking Network (RMN)

The **Residual Masking Network (RMN)** [15] is a state-of-the-art deep learning model designed for facial expression recognition, utilizing **Residual Masking Blocks** to enhance the extraction of critical multi-scale facial features. This innovative architecture processes features hierarchically, emphasizing the most relevant regions of the face while suppressing irrelevant or noisy data. The model concludes with a softmax layer to classify facial expressions into seven distinct categories: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral.

Fine-tuned on the **FER2013 dataset**, the RMN achieves an impressive accuracy of **74.14%**, outperforming several widely recognized architectures such as VGG19 and ResNet34. Its advanced feature masking mechanism ensures robust performance even under challenging conditions, such as low-resolution images or diverse facial poses.

RMN's design makes it both efficient and versatile, supporting applications in static image analysis and real-time video emotion recognition. Pre-trained weights are readily available, allowing seamless integration into various projects without the need for extensive training. The model is particularly well-suited for use in fields such as human-computer interaction, adaptive interfaces, and behavioral analysis, where accurate and efficient emotion recognition is crucial.

4.2.5. EmoNet

EmoNet [16] is a sophisticated model designed for both discrete and continuous emotion recognition, capable of predicting categorical emotions alongside valence and arousal levels. The model offers two configurations: one for **5 emotional classes** and another for **8 classes**, with accuracies of **82%** and **75%** respectively when evaluated on the AffectNet dataset. Additionally, EmoNet predicts facial landmarks, enriching its analysis capabilities for real-world applications.

Trained on the **AffectNet dataset**, EmoNet is optimized for naturalistic conditions, where facial expressions may vary significantly due to environmental factors or individual differences. The model's predictions include: - **Discrete Emotions:** Categories such as Neutral, Happy, Sad, Surprise, Fear, Disgust, Anger, and Contempt (for the 8-class model). - **Valence and Arousal:** Continuous measures indicating emotional intensity and activation levels, respectively.

The model employs the **SFD detector** from the face-alignment repository for facial feature localization. Although this detector ensures high accuracy, it operates at relatively slower speeds compared to other detection methods. EmoNet's architecture integrates deep learning techniques to achieve robust performance across diverse facial expressions, making it suitable for applications in affective computing and emotion-aware systems.

EmoNet provides tools for analyzing static images and video streams, supporting dynamic emotion recognition in real-time scenarios. Its dual focus on discrete and continuous emotion analysis allows for a comprehensive understanding of human affect, enabling its use in fields such as psychology, human-computer interaction, and entertainment.

4.3. Summary of Models

Model	Dataset	Number of Classes	Valence/Arousal	Accuracy
DeepFace	FER2013	7	No	57.42%
HSEmotionONNX (16 Mb)	AffectNet	7	No	64.57%
HSEmotionONNX (16 Mb)	AffectNet	8	No	61.32%
HSEmotionONNX (30 Mb)	AffectNet	7	No	66.29%
HSEmotionONNX (30 Mb)	AffectNet	8	No	63.03%
Vision Transformer	FER2013	7	No	71.16%
Residual Masking Network	FER2013	7	No	74.14%
EmoNet	AffectNet	5	Yes	82%
EmoNet	AffectNet	8	Yes	75%

5. Prototype and Demo Set-up

This section aims at providing a detailed overview of the prototype setup and configuration for the MoodPilot project. It includes the hardware specifications, software environment, and deployment strategies for the Facial Emotion Recognition (FER) models on edge devices. The Raspberry Pi 3B+ serves as the primary edge device for running the FER models, with a focus on real-time performance and resource efficiency. Moreover, an overview over the data collection system and the frontend and backend components is provided, showcasing the system's architecture and functionality.

5.1. Raspberry Pi 3B+ Configuration

The edge device used during the prototype phase to run the FD (Face Detection) and FER (Facial Expression Recognition) models is the Raspberry Pi 3B+. Below, the hardware specifications and setup details are provided.

5.1.1. Raspberry Pi 3 Model B+

The Raspberry Pi 3 Model B+ is a third-generation single-board computer, featuring:

- **Processor**: 1.4GHz 64-bit quad-core Broadcom BCM2837B0, Cortex-A53 (ARMv8) SoC.
- **Memory**: 1GB LPDDR2 SDRAM.
- **Wireless Connectivity**: Dual-band 2.4GHz and 5GHz IEEE 802.11.b/g/n/ac wireless LAN, Bluetooth 4.2/BLE.
- **Ethernet**: Gigabit Ethernet over USB 2.0, with a maximum throughput of 300 Mbps.
- **GPIO**: Extended 40-pin GPIO header.
- **Camera Support**: CSI camera port for connecting a Raspberry Pi camera.
- **Power Input**: 5V/2.5A DC power input.

5.1.2. Camera Module

For real-time detection, the Raspberry Pi Camera Module 2 (v2.1) was used. The module includes:

- **Sensor**: Sony IMX219, 8-megapixel sensor.
- **Capabilities**:
  - High-definition video recording: 1080p30, 720p60, VGA90 video modes.
  - Still image capture.
- **Connection**: Attaches via a 15cm ribbon cable to the CSI port on the Raspberry Pi.
- **Software Support**: Numerous third-party libraries are available, including the Picamera Python library. Refer to the "Getting Started with Picamera" resource for additional guidance.

5.1.3. Operating System Setup

The operating system was installed on a 128GB SanDisk Ultra microSD card with speeds up to 140MB/s. The Raspberry Pi Imager was used to install the following version:

- **OS Version**: Raspberry Pi OS with desktop.
- **Release Date**: November 19th, 2024.
- **System**: 64-bit.
- **Kernel Version**: 6.6.

- **Debian Version**: 12 (Bookworm).
- **Image Size**: 1,179MB.

To verify system details, use the command:

```
uname -a
```

Sample output:

```
Linux raspberrypi 6.6.62+rpt-rpi-v8 #1 SMP PREEMPT Debian 1:6.6.62-1+rpt1
↪  (2024-11-25) aarch64 GNU/Linux
```

5.1.4. Python Environment and Dependencies

The installed Python version is 3.11.2. To deploy the system, all required dependencies can be installed automatically using the following command from the root of the project:

```
pip install -r requirements_raspberrypi.txt
```

For testing models on a laptop, a separate `requirements.txt` file is available to install the necessary dependencies.

5.2. Raspberry Pi 3B+ Demo

For the deployment on the Raspberry Pi, which differs from deployment on a basic notebook used for testing purposes, adjustments were required due to compatibility issues. On the notebook, OpenCV was utilized for both camera input and frame processing. However, with the Raspberry Pi 3B+ running the latest Raspberry Pi OS, OpenCV alone could not interface with the Picamera module. Additionally, the older Python Picamera library was incompatible with the newer OS version. After extensive testing and troubleshooting, the following configuration was successfully implemented.

5.2.1. Camera Integration

The new `picamera2` library was employed for capturing frames from the Picamera module. Using the `Picamera2` class, an instance was created to capture a frame, which was subsequently processed using OpenCV for resizing and overlaying information for real-time preview. To verify the camera configuration with real-time preview, execute the following command from the root of the project:

```
python3 Prototype/FER/camera.py
```

This command opens an OpenCV window titled "Camera Preview," displaying real-time frames captured by the camera, as shown in Figure 2. To exit the preview, press "q" while the window is focused or simply close the window.

5.2.2. Emotion Detection Deployment

For the emotion detection system on the Raspberry Pi 3B+, the "HSEmotion_Onnx" FER model was selected, paired with HAAR Cascade for face detection. This decision was based on the limited computational power of the Raspberry Pi 3B+ and the unavailability of a Neural Compute Stick 2 (NCS2) accelerator. The specific implementation is in the `Prototype/FER/Models/HSEmotionONNX/hsemotion_onnx.py` file.

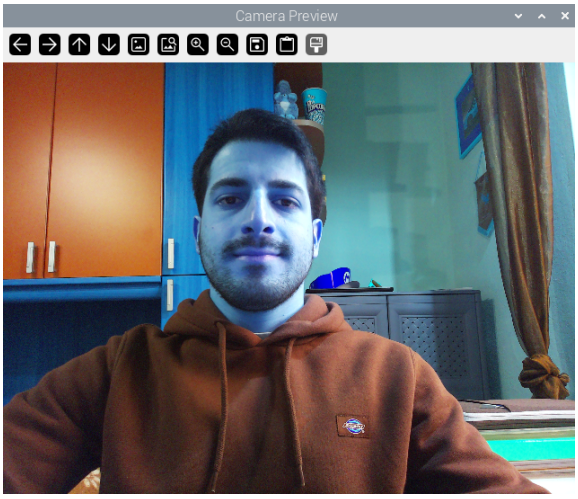


Figure 2: Real-time camera preview using the Picamera module on the Raspberry Pi 3B+.

To explore the script's options, use:

```
cd Prototype/FER/Models/HSEmotionONNX/
```

```
python3 hsemotion_onnx.py --help
```

5.2.2.1. Usage

```
usage: hsemotion_onnx.py [-h] [--video VIDEO] [--no-preview]

HSEmotionONNX: Real-time facial emotion recognition.

options:
  -h, --help     show this help message and exit
  --video VIDEO  Path to video file or 'camera' for live feed.
  --no-preview   Disable video preview during processing.
```

- **Video Input**: By default, or when `--video camera` is specified, the Picamera feed serves as input. Alternatively, a stored video can be provided.
- **Preview Mode**: The `--no-preview` option disables the real-time video preview to save FPS, aligning with the final deployment scenario where passengers will not view the preview during the trip.

To run the model with the default settings:

```
python3 hsemotion_onnx.py
```

5.2.3. Output Examples

The camera feed and preview window display real-time emotion detection. Sample outputs include:

- Figure 3a
- Figure 3b
- Figure 3c
- Figure 3d

Each image illustrates successful detection of respective emotions (neutral, happiness, surprise, and disgust). To exit, press "q" or use `CTRL+C`. Graceful termination is implemented via a signal handler for SIGINT, ensuring proper program termination.

5.2.4. Performance Metrics

The FPS during execution highlights the computational limitations of the Raspberry Pi 3B+:

- **With Preview**: 0.50 FPS (Figure 4a)
- **Without Preview**: 0.87 FPS (Figure 4b)

Modifica in modo tale che le imagini siano affiancate a coppie

5.2.5. Logging

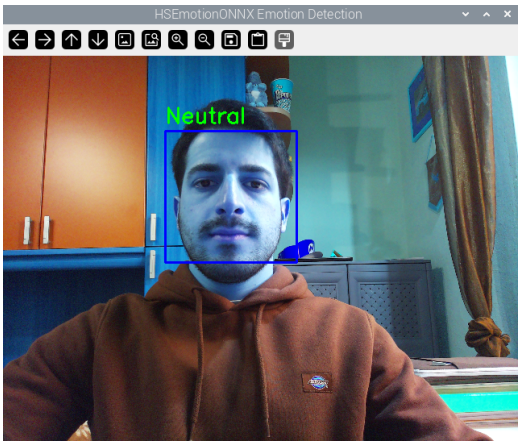
All detected emotions, along with timestamps, are logged in the

`Prototype/FER/Models/HSEmotionONNX/logs/`

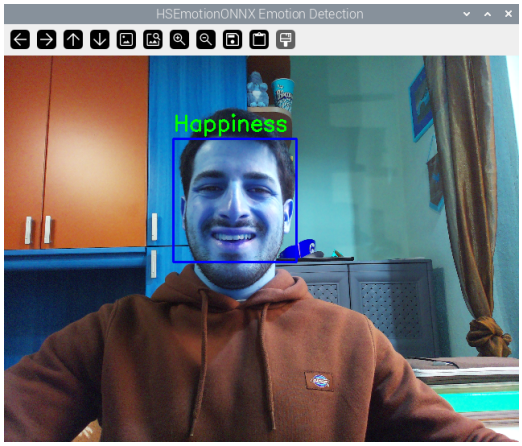
folder. A unique log file is created for each execution, with average FPS appended at the end.

5.2.6. Recommendations

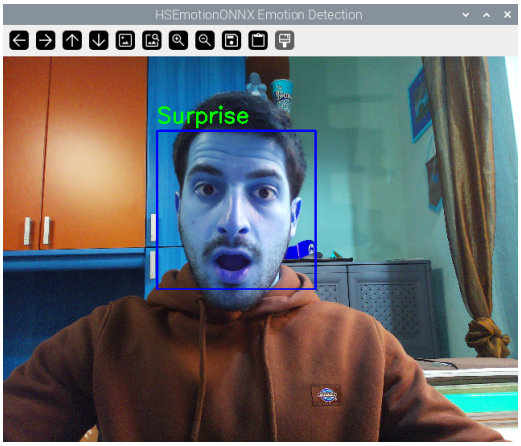
Although the processed FPS is low (more than one second per frame), the successful deployment demonstrates the capability of integrating modern Python (3.11) and software tools on the dated Raspberry Pi 3B+. For practical applications, upgrading to a more powerful edge device is recommended to enhance performance. Future work includes testing this deployment on newer Raspberry Pi models to evaluate potential improvements in real-time emotion detection, enabling the "MoodPilot" system to assess passenger moods during trips.



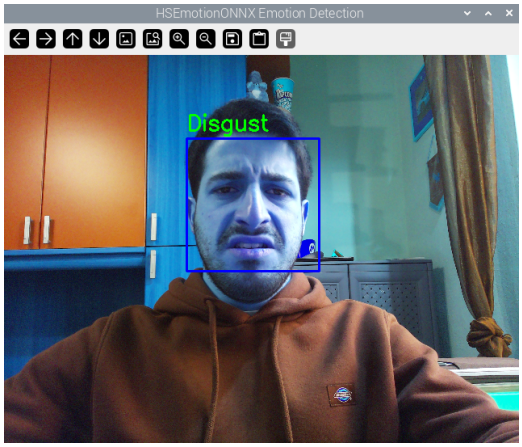
(a) Detection of a neutral facial expression.



(b) Detection of a happy facial expression.

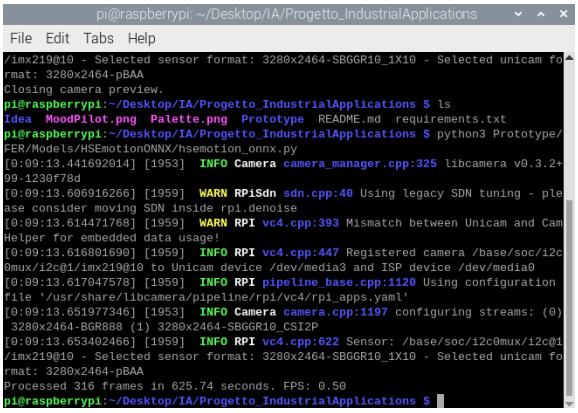


(c) Detection of a surprised facial expression.

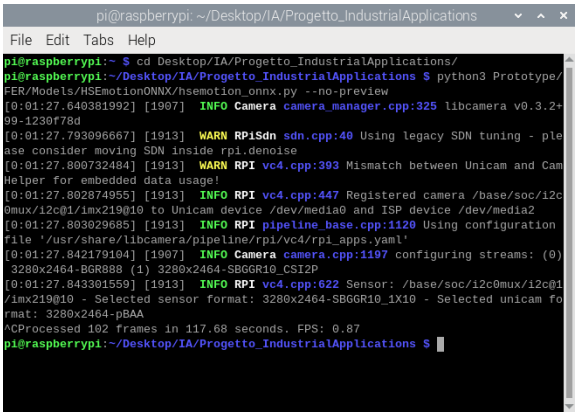


(d) Detection of a disgusted facial expression.

Figure 3: Real-time emotion detection using the HSEmotionONNX model. The figures depict successful detection of neutral, happiness, surprise, and disgust emotions.



(a) FPS during emotion detection with video preview enabled.



(b) FPS during emotion detection without video preview.

Figure 4: Frames per second (FPS) during emotion detection using the HSEmotionONNX model on the Raspberry Pi 3B+.

5.3. Data Collection System

6. Conclusion

References

[1] W. contributors, "Emotion classification — Wikipedia, the free encyclopedia." 2023. Available: https://en.wikipedia.org/wiki/Emotion_classification

[2] T. F. Murphy, "Circumplex model of arousal and valence." 2024. Available: https://psychology fanatic.com/circumplex-model-of-arousal-and-valence/

[3] M. Basili and M. A. Rossi, "Platform-mediated reputation systems in the sharing economy and incentives to provide service quality: The case of ridesharing services," *Electronic Commerce Research and Applications*, 2019, doi: 10.1016/j.elerap.2019.100835.

[4] L. Mitropoulos, A. Kortsari, and G. Ayfantopoulou, "A systematic literature review of ride-sharing platforms, user factors, and barriers," *European Transport Research Review*, vol. 13, p. 61, 2021, doi: 10.1186/s12544-021-00522-1.

[5] L. Mitropoulos, "Understanding ride-sharing systems in urban areas: The role of location, users and barriers." 2020.

[6] S. Zepf, J. Hernandez, A. Schmitt, W. Minker, and R. W. Picard, "Driver emotion recognition for intelligent vehicles: A survey," *ACM Computing Surveys*, vol. 53, no. 3, pp. 64:1–64:30, 2020, doi: 10.1145/3388790.

[7] V. Karas, D. M. Schuller, and B. W. Schuller, "Audiovisual affect recognition for autonomous vehicles: Applications and future agendas," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 6, pp. 4918–4932, 2024, doi: 10.1109/TITS.2023.3333749.

[8] W. Li, G. Li, R. Tan, *et al.*, "Review and perspectives on human emotion for connected automated vehicles," *Automotive Innovation*, vol. 7, pp. 4–44, 2024, doi: 10.1007/s42154-023-00270-z.

[9] M. G. C. A. Cimino, A. Di Tecco, P. Foglia, and C. A. Prete, "Using emotion recognition and temporary mobile social network in on-board services for car passengers," in *Smart cities, green technologies, and intelligent transport systems*, C. Klein, M. Jarke, J. Ploeg, M. Helfert, K. Berns, and O. Gusikhin, Eds., Cham: Springer Nature Switzerland, 2023, pp. 158–171.

[10] M. Sambare, "FER-2013: Learn facial expressions from an image." https://www.kaggle.com/datasets/msambare/fer2013, 2020.

[11] M. H. Mahoor, "AffectNet: Annotated facial database for affective computing." http://mohammadmahoor.com/affectnet/.

[12] S. I. Serengil, "DeepFace: A lightweight face recognition and facial attribute analysis library for python." https://github.com/serengil/deepface, 2023.

[13] A. Savchenko, "Facial expression recognition with adaptive frame rate based on multiple testing correction," in *Proceedings of the 40th international conference on machine learning (ICML)*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., in Proceedings of machine learning research, vol. 202. PMLR, 2023, pp. 30119–30129. Available: https://proceedings.mlr.press/v202/savchenko23a.html

[14] Todor Pakov, "Vit-face-expression (revision 78ed8d3)." Hugging Face, 2024. doi: 10.57967/hf/2289.

[15] L. Pham, T. H. Vu, and T. A. Tran, "Facial expression recognition using residual masking network," in *2020 25th international conference on pattern recognition (ICPR)*, IEEE, 2021, pp. 4513–4519.

[16] A. Toisoul, J. Kossaifi, A. Bulat, G. Tzimiropoulos, and M. Pantic, "Estimation of continuous valence and arousal levels from faces in naturalistic conditions," *Nature Machine Intelligence*, 2021, Available: https://www.nature.com/articles/s42256-020-00280-0