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).

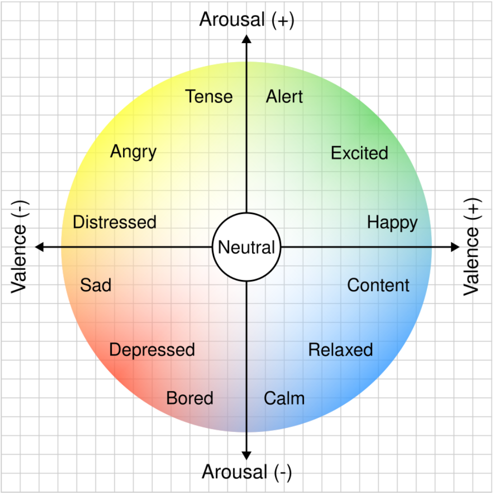

Figure 1: The Circumplex Model of Emotion maps emotions within a two-dimensional space defined by valence (X-axis) and arousal (Y-axis). Positive and negative valence correspond to pleasant and unpleasant emotions, respectively, while arousal indicates emotional intensity. Image taken from en.wikipedia.org/wiki/Emotion_classification

- **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

2.1. Bidirectional Feedback Limitations in Ride-Hailing Services

2.2. Facial Emotion Recognition (FER) in Automotive Contexts

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" [3], 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

4. Facial Expression Recognition (FER) Performance Analysis

4.1. Datasets

The datasets described here will be referenced later in the Models section to indicate their use in training and evaluating the models. Below is a summary of the datasets:

4.1.1. FER2013

[4] is a dataset comprising 32,298 grayscale images of faces, each sized 48x48 pixels, labeled with seven emotions: Angry (0), Disgust (1), Fear (2), Happy (3), Sad (4), Surprise (5), and Neutral (6). Images are centered and uniformly scaled. The dataset includes 28,709 training samples and 3,589 test samples, with the task being to classify facial expressions into one of the seven categories.

4.1.2. AffectNet

[5] addresses the scarcity of annotated facial expression datasets in the wild, particularly for the **continuous dimensional model** (e.g., valence and arousal) alongside **discrete emotions** (categorical model). It contains over 1 million facial images collected from the internet using 1,250 emotion-related keywords in six languages. Approximately 440,000 images were manually annotated for eight discrete emotions and valence/arousal intensity. The dataset also includes “contempt” as an eighth emotion, making it more comprehensive than FER2013. AffectNet facilitates research in both categorical and dimensional emotion models.

4.2. Models

This section introduces several models for Facial Expression Recognition (FER). Each model is presented with a brief description, the dataset it utilizes (if not otherwise specified, for both training and testing), and its performance metrics on that dataset.

4.2.1. DeepFace

Description: DeepFace [6], [7] is a lightweight Python framework for face recognition and facial attribute analysis, including emotion detection, age, gender, and ethnicity prediction. It integrates various state-of-the-art models such as VGG-Face, FaceNet, and ArcFace.
Dataset Used: FER2013
Accuracy: 57.42%

4.2.2. HSEmotionONNX

Description: [8], [9], [10], [11], [12] A collection of ONNX-compatible (Open Neural Network Exchange standard) models *pre-trained* for face identification using VGGFace2 (Dataset for Face Recognition) and optimized for emotion recognition on AffectNet.
Dataset Used for Fine-Tuning: AffectNet
Accuracy:

- enet_b0_8_best_vgaf.pt: 61.32% (8 classes), 64.57% (7 classes)

- enet_b2_8.pt: 63.03% (8 classes), 66.29% (7 classes)

Inference Times:

- enet_b0_8_best_vgaf.pt: 59 ± 26 ms
- enet_b2_8.pt: 191 ± 18 ms

Model Sizes:

- enet_b0_8_best_vgaf.pt: 16 MB
- enet_b2_8.pt: 30 MB

The enet_b0_8_best_vgaf model is preferred due to its faster inference time, achieving higher FPS with minimal accuracy trade-offs.

4.2.3. Vision Transformer (ViT) for Facial Expression Recognition

Description: [13] A Vision Transformer model fine-tuned (from vit-base-patch16-224-in21k) for emotion recognition on facial images.
Dataset Used for Fine-Tuning: FER2013
Accuracy:

- Validation Set: 71.13%
- Test Set: 71.16%

4.2.4. Residual Masking Network (RMN)

Description: RMN [14] leverages Residual Masking Blocks to process facial features across multiple scales, ending with a 7-class softmax for emotion classification.
Dataset Used: FER2013
Accuracy: 74.14%

4.2.5. EmoNet

Description: [15] Trained models available for 5 and 8 emotional classes, also predicting valence, arousal, and facial landmarks.
Dataset Used: AffectNet

Accuracy:

- 5 Classes: 82%
- 8 Classes: 75%

Face Detection: Utilizes the SFD detector from the face-alignment repository, noted for its high accuracy but slower performance.

5. Prototype and Demo Set-up

5.1. Raspberry Pi 3B+ Configuration

5.2. 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://psychologyfanatic.com/circumplex-model-of-arousal-and-valence/

[3] 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.

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

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

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

[7] S. I. Serengil and A. Ozpinar, “HyperExtended LightFace: A facial attribute analysis framework,” in *2021 international conference on engineering and emerging technologies (ICEET)*, IEEE, 2021, pp. 1–4. doi: 10.1109/ICEET53442.2021.9659697.

[8] 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

[9] A. V. Savchenko, “Facial expression and attributes recognition based on multi-task learning of lightweight neural networks,” in *Proceedings of the 19th international symposium on intelligent systems and informatics (SISY)*, IEEE, 2021, pp. 119–124. Available: https://arxiv.org/abs/2103.17107

[10] A. V. Savchenko, “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 (CVPR) workshops*, 2022, pp. 2359–2366. Available: https://arxiv.org/abs/2103.17107

[11] A. V. Savchenko, “MT-EmotiEffNet for multi-task human affective behavior analysis and learning from synthetic data,” in *Proceedings of the european conference on computer vision (ECCV 2022) workshops*, Springer, 2023, pp. 45–59. Available: https://arxiv.org/abs/2207.09508

[12] A. V. Savchenko, L. V. Savchenko, and I. Makarov, “Classifying emotions and engagement in online learning based on a single facial expression recognition neural network,” *IEEE Transactions on Affective Computing*, 2022, Available: https://ieeexplore.ieee.org/document/9815154

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

[14] 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.

[15] 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