User State Detection

Exploring State-of-the-Art Tools for Emotion Detection

Birol Arin[1 [7406390]] and Julian Stricker[1 [7380409]]

[1]Department of Information Systems for Sustainable Society Faculty of Management, Economics and Social Sciences, University of Cologne
barin@smail.uni-koeln.de
jstricke@smail.uni-koeln.de

Abstract. This paper focuses on the current state of tools and frameworks for automatically recognizing user states, focusing on emotion recognition from speech and video data Through a structured review and comparison, eight tools were selected based on their architecture, modality, impact and performance. The evaluation highlights advantages and disadvantages of the different approaches. CNN-based models, transformer-based architectures and multimodal systems were analyzed. An implementation with the LibreFace tool was also used to demonstrate the practicality of real-time emotion recognition. The results highlight the importance of balancing accuracy, computational efficiency, and usability in practical applications. Accordingly, future research should focus on improving multimodal integration and continue to pursue new technologies such as transformer-based architectures.

Keywords: affective computing · facial emotion recognition · deep learning · machine learning

1 Introduction

The recognition and analysis of emotions has gained considerable importance in recent years, particularly in the context of human-computer interaction. Emotions play a central role in interpersonal relationships, in cognition, in perception and in many other aspects of life [2]. Understanding emotions has become essential for the daily functioning of humans, as the acquisition and experience of emotions is fundamental to communication in social environments [2].

Automatic emotion recognition from facial expressions or speech, has gained significant momentum with the success of affective computing (AC) and cognitive computing [23]. Researchers use different modalities to study emotion recognition, including speech, text, facial features and EEG-based brainwaves [2]. This multimodal approach reflects the complexity of human perception, which includes visual, auditory and other sensory information [7]. The development of emotion sensitive interfaces is taking place in a variety of domains, including gaming, mental health and learning technologies. The basic idea behind most

AC systems is that automatically recognizing and responding to a user's emotional states while interacting with a computer can improve the quality of the interaction and make the computer interface more user friendly, enjoyable and efficient [10]. Facial expressions play a significant role in recognizing emotions, as they are considered a universal language that transcends cultural and ethnic differences. Facial expressions can provide information about a person's mental state that is otherwise difficult to recognize [33].

Research in the field of face recognition can be divided into feature-based and holistic approaches. The early work on face recognition was feature-based and attempted to define a low-dimensional representation of faces explicitly based on distance, area and angle relationships [4]. However, with the increasing prevalence of cameras in the Internet of Things (IoT) and the ability to use facial recognition technologies for context enhancement, a large gap has opened up between publicly available facial recognition systems and the most advanced private systems [4]. Given the increasing and diverse use of human-computer interaction, emotion recognition technologies offer an opportunity to promote pleasant or intuitive user interactions [2]. In recent years, interest in machine learning and recognition of affective and cognitive mental states as well as interpretation of social signals, especially based on facial expressions and general facial behavior, has increased [8]. These developments underline the importance of emotion recognition in various application areas, from improving the customer experience in e-commerce, helping in learning environments, assistance in aviation and aerospace to supporting psychological diagnostics and therapy [7, 19, 22, 32, 41]. In the field of human-machine interaction, emotion recognition enables more natural and intuitive communication between humans and machines, which can lead to more effective and pleasant interactions [35].

2 Theoretical Background

2.1 Emotions according to Ekman

Paul Ekman is considered one of the most well known psychologists in the field of emotion research [34]. In collaboration with Friesen (1978), they jointly developed the Facial Action Coding System (FACS), which is still an important tool for researching facial expressions and emotions [16]. FACS is a system to describe any visible facial movement [15, 16, 37]. In contrast to the categorization of emotions, the goal of FACS is to identify and encode basic building blocks of facial expressions known as action units (AU) [8, 16, 37]. FACS is used in several research areas as a basis for the objective measurement of facial expressions and their correlation with emotions [8, 10, 14, 33, 37].

Ekman (1992) developed a widespread theory according to which certain emotions are universal and biologically anchored [15, 24]. According to Ekman (1992), these emotions have developed over the course of evolution to fulfill

adaptive functions in coping with basic life tasks [15, 24]. Ekman classified the emotions happiness, sadness, fear, anger, disgust and surprise (and neutral) as basic emotions (Figure 1) [2, 3, 15, 24, 25] and argued that certain facial expressions for basic emotions can be identified across cultures making them universal [15]. He also claimed that each basic emotion not only has a unique facial expression, but is also associated with a specific pattern of changes in the autonomic nervous system [15].

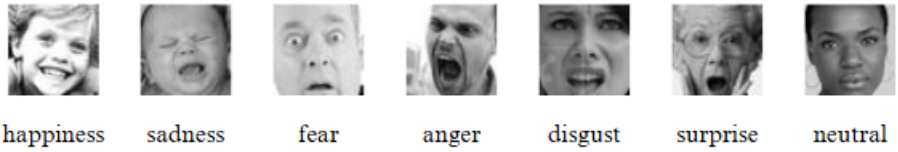

Fig. 1. Basic emotions according to Ekman (1992) from FER-2013 dataset [26]

Although Ekman's theory of basic emotions is influential and widely spread, the assumption of cultural universality has been largely refuted. Studies show that the expression and interpretation of emotions are influenced by cultural norms and learning processes [24, 25]. In fact are emotions more dynamic and cannot be divided into static categories, as they are much more context dependent processes [25, 43]. However, these basic emotions are still applied in the field of facial emotion recognition, which is why they are also defined as basic emotions in the following paper.

2.2 Emotions according to Russel

The valence-arousal model offers a dimensional approach to understanding emotions, in contrast to categorical models such as those proposed by Ekman. According to Russell , emotions are best represented within a two-dimensional circumplex model comprising the valence (pleasure-displeasure) and arousal (activation-deactivation) axes [36]. This framework captures the dynamic and continuous nature of emotional states, allowing for a nuanced representation of intensity and similarity between emotions [21]. Although arousal correlates with physiological activation, such as heart rate and electrodermal activity, valence often reflects subjective pleasantness and is related to facial expressions [9]. The model's dimensional nature supports its application in diverse contexts, including emotion recognition systems, where it facilitates the integration of multimodal data for detecting subtle and blended emotional states [9, 21]. A unique limitation of the valence-arousal model is its reliance on the assumption of independence between valence and arousal dimensions, despite evidence suggesting that these dimensions are often correlated, complicating their separate measurement and interpretation [21].

2.3 Facial Emotion Recognition

Facial Expression Recognition or Facial Emotion Recognition (FER) is a computer vision task whose goal is to analyze static images and videos of facial expressions to identify the emotions felt by a person [1, 28]. Due to the rapid development in the research fields of machine learning, deep learning and artificial intelligence as well as the prevalence of IoT and sensors FER has become increasingly significant in recent years [1, 28, 29, 38, 39]. FER consists of the main steps of face recognition including all components of the face, feature extraction and lastly emotion classification [1, 28]. These steps are often implemented in modern FER systems using advanced deep learning methods such as Convolutional Neural Networks (CNN) [28, 35]. These technologies play an important role in human-computer interaction and can enable computers to recognize and respond to human emotions [28, 33]. From driving safety to entertainment and healthcare to aerospace, FER can be applied in diverse areas of our day to day life [1, 29].

3 Method

3.1 Method

This paper followed an iterative search process across multiple databases, including Google Scholar, IEEE Xplore, ResearchGate, GitHub, and arXiv. The authors used search terms focused on facial expression recognition, deep learning, open-source tools, and real-time processing to identify relevant studies.

The selection criteria considered both pre and post 2020 research. Older studies were included based on their influence, citation count and peer-reviewed status, while newer research was selected based on its novelty, power and emerging architectures such as transformers. The authors have incrementally refined the study and paper selection based on relevance to user state recognition and FER, the availability of open source implementations, and the diversity of model architectures.

A total of 46 articles and papers were reviewed, leading to the selection of eight tools or models covering a wide range of modalities and model architectures. These tools were evaluated in terms of their applicability in emotion recognition and their potential for real-time recognition of user states.

3.2 Research Objective

The central aim of this paper is to compare unimodal and multimodal emotion recognition systems to evaluate their suitability and feasability for real life application scenarios, taking into account performance and implementation effort. Through a structured comparison of established and state-of-the-art tools and

models, strengths and weaknesses in terms of architecture, accuracy and real-time capability will be identified. The analysis of eight models examines how factors such as open source adaptability, documentation quality or multimodality influence practical usability in real-time systems.

A central contribution of this paper is to combine theoretical findings with practical validation: A comparative tabular analysis of the different models provides an overview and categorization based on various characteristics and criteria. Additionally a prototype implementation demonstrates how a selected tool perform under controlled conditions, highlighting trade-offs between computational efficiency and classification robustness. The analysis prioritizes open-source solutions to provide reproducibility in an academic context while addressing gaps in cross-domain integration which is a critical aspect for future systems that will combine facial, speech and physiological data streams for a multimodal approach.

The aim of the study is to create an evidence-based basis for the selection of FER technologies that takes into account both academic requirements and industrial scalability.

4 Results

4.1 Overview of Models and Tools

In this section, eight selected models and tools are examined in more detail. This includes both established and modern state-of-the-art technologies. The selected technologies include four feature extraction tools enabling emotion recognition and four emotion recognition tools for direct classifications:

openSMILE openSMILE, short for *Open Speech and Music Interpretation by Large-Space Extraction*, is a versatile and open-source toolkit designed for audio feature extraction, widely applied in fields such as speech and music analysis, paralinguistics, and emotion recognition. Developed to support real-time and offline processing, openSMILE provides a modular architecture capable of extracting thousands of features, including low-level descriptors (LLDs) such as pitch, loudness, Mel-Frequency Cepstral Coefficients (MFCCs), and spectral features. Its key contribution lies in its ability to provide a standardized and comprehensive feature set that is compatible with machine learning frameworks, enabling high accuracy in tasks such as automatic speech recognition and emotion detection [17, 18].

openSMILE is extensively used for emotion recognition by extracting acoustic features that correlate with emotional states. For instance, in a real-time emotion detection framework, openSMILE extracted LLDs, including pitch and energy, which were then encoded into fixed-size embeddings for processing by a

recurrent neural network (RNN). This method achieved high accuracy in classifying emotions like happiness and anger in conversational datasets [11]. Similarly, in a multi-layer perceptron (MLP) classifier approach, openSMILE provided a robust set of 998 features, including intensity, spectral flux, and MFCCs. These features were key to achieving over 83% accuracy when combined with data augmentation techniques on the Berlin Emotional Speech database [46]. Additionally, openSMILE's integration into a hybrid framework with GPT-3.5 for emotion detection combined semantic analysis with acoustic features, yielding a balanced and effective recognition system [44].

Despite its versatility and extensive feature set, openSMILE has certain limitations. Configuring its feature extraction pipeline requires domain expertise, and real-time processing of large datasets can be computationally demanding. Moreover, its performance may vary with audio quality, environmental noise, and cultural differences in emotional expression [30, 44]. These challenges underscore the need for continued development to enhance openSMILE's robustness and scalability in real-world applications.

Openface OpenFace is the first open source facial behavior analysis tool that takes into consideration different modalities of facial behavior. The tool was developed by researchers at the University of Cambridge and implements multiple algorithms for the recognition of facial landmark motion, head pose (orientation and motion), facial expressions and eye gaze, which play an important role in individual and social human behavior and interactions [8]. Since most state-of-the-art technologies are dominated by industry and government datasets and often lack information in published scientific papers, re-implementation was almost impossible. Either values for hyperparameters, data normalization and cleaning processes or exact training protocols, which are important for the development of systems with real-world data, were missing [4, 8]. Consequently, the development of OpenFace was driven by the necessity of considering various modalities. Furthermore, the system was designed to operate in real time.

First, the system recognizes faces in images or videos. It then processes each recognized face to generate a normalized input of fixed size for the neural network [4]. This neural network acts as a feature extractor and generates a low-dimensional representation that characterizes a person's face. This compact representation is crucial for efficient use in classification or clustering methods [4]. OpenFace defines facial behavior as a combination of movements of the facial landmarks, head posture (orientation and movement), facial expression and gaze direction [8]. The system uses Conditional Local Neural Fields (CLNF) to recognize and track facial features, which are an instance of a Constrained Local Model (CLM). The model groups 68 facial landmarks and is then validated with a CNN [8]. Furthermore it is capable of identifying both the presence and intensity of AU's. As a result, it is possible to analyze their occurrence, coexistence and dynamics [8]. In addition, head pose and gestures play an important role in

the perception and expression of emotions and social signals. The direction of gaze is also important in the assessment of attention, social skills, mental health and the intensity of emotions [8].

OpenFace is characterized by its real-time capability and does not rely on a GPU for processing. The system offers a variety of usage options, including a graphical user interface, a command line and a real-time messaging system. It is able to process real-time video streams from webcams, recorded video files, image sequences and individual images and provides the model training code [8].

OpenFace 2.0 The OpenFace researchers have expanded the tool and released OpenFace 2.0 in 2018 [7]. This tool is also primarily aimed at providing support for research in the areas of computer vision and machine learning, AC and interactive applications based on facial behavior analysis. The extension improves the facial landmark detection, head pose estimation, facial AU recognition and eye gaze estimation and can therefore better identify and predict them [7]. While OpenFace uses older algorithms such as the dlib-based face detector and has problems with non-frontal faces and masking [8], OpenFace 2.0 uses a Convolutional Experts Constrained Local Model (CE-CLM), which works more precisely and faster thanks to optimizations such as model simplification and intelligent multiple hypotheses. As a result, masked faces and difficult lighting conditions can be processed much better [7]. 3D modeling keeps the head pose of the projection robust, and an optimized CLNF is used to detect detailed eyelids, irises and pupils [7]. While OpenFace 1.0 had difficulties to reliably detect AU intensities in natural videos [8], the extension integrates improved methods, such as person-specific normalization and predictive correction, allowing for higher accuracy. OpenFace 2.0 is an extension and improvement of the previous tool and offers more detailed documentation and new implementation options, such as cross-platform support and additional scripts that help with the extraction, reading and visualization of facial features [7].

LibreFace LibreFace is an open source tool for analyzing facial expressions that was published by researchers from the University of Southern California in 2024 which was presented at the Winter Conference on Applications of Computer Vision (WACV) [12]. The tool offers a comprehensive solution for FER tasks and can also perform them in real time. It uses deep learning models for different subtasks, including the recognition of AU's according to Ekman (1972), the recognition of landmarks and the classification of emotions [12]. Using MediaPipe (a cross-platform that provides customizable machine learning solutions for livestreaming) [20], a face mesh is created to locate landmark facial features such as eyes, eyebrows, nose and mouth [12]. The facial images are then aligned with the previously recognized landmarks to ensure consistent and correct positioning. While other systems only use CNNs RNNs or Transformers, LibreFace uses a pre-trained Masked Auto-encoder (MAE) to extract facial features and their features. The extracted AU and their intensity are then processed through

linear regression layers and classification layers to predict the labels. The model uses feature wise distillation, which transfers the output of the MAE model into a ResNet-18 model, increasing inference efficiency [12].

LibreFace and its models were pre-trained on widely spread datasets such as ImageNet, EmotioNet, AffectNet, DISFA and FFHQ, which improved the performance of the model. While LibreFace has an average Pearson correlation for AUs on the DISFA dataset, it achieves state of the art performance in emotion classification on the AffectNet and RAF-DB datasets in comparison to other models [12].

The open-source toolkit provides a real-time solution for emotion recognition and classification, either purely CPU based or with GPU accelerated versions. It represents a simpler, free-to-use solution that differs from most research oriented solutions. The entire system is cross-platform, source code is available in C# or Python as open source, components are developed as .NET libraries and it provides an intuitive GUI [12].

DeepFace DeepFace is a deep learning-based face recognition system developed by Facebook AI Research in 2014 to achieve near-human accuracy in facial verification tasks. It introduced a novel approach that combines 3D face alignment and a deep CNN architecture to extract and classify facial features [31, 42]. A key innovation was the use of Locally Connected Layers (LCL), which, unlike traditional CNNs, do not share weights across spatial regions. This design allowed DeepFace to focus on region-specific features like the eyes and mouth, which in result improved the accuracy of face verification [31]. The system was trained on a dataset containing over 4 million labeled face images from approximately 4,000 identities, achieving a 97.35% accuracy on the Labeled Faces in the Wild (LFW) benchmark [42].

Although DeepFace was originally developed for face verification, it has been effectively adapted for emotion recognition by leveraging its pre-trained deep learning models and fine-tuning them on emotional datasets. The methodology involves a four-step process: face detection, alignment, feature extraction, and emotion classification. The model begins with face detection and 3D alignment, ensuring standardized face positioning. Next, the CNN extracts numerical feature embeddings from the face, which are then used to classify emotional expressions using models like VGG-Face, ArcFace, and FaceNet. These models can detect seven basic emotions: happiness, sadness, fear, anger, disgust, surprise, and neutrality [5, 27]. Real-time emotion detection has been demonstrated using datasets like CK+ and FER+, with accuracy rates reaching 90,25% in controlled environments [27, 45].

DeepFace has shown promising results in emotion recognition tasks, achieving accuracy rates as high as 95.93% on normalized training data and 92.02% on

testing data when trained on the FER+ dataset [45]. Real-time applications were tested with a webcam setup that enables live emotion detection with successful classification of primary emotions like happiness, sadness, and anger. Potential applications include healthcare (for emotional health monitoring), driver safety systems (for detecting stress or fatigue), security and surveillance, and human-computer interaction [5].

Despite its strengths, DeepFace has several limitations when applied to emotion recognition. One major problem is the data imbalance in emotional datasets, where some emotions are underrepresented, leading to reduced generalization. Furthermore, lighting conditions, occlusions (e.g. glasses, masks), and cultural variations in emotional expressions can affect accuracy [27, 45]. The computational complexity of LCL also makes real-time processing challenging. Modern approaches using standard CNNs with batch normalization have proven to be more efficient in recent years [31, 42].

ResEmoteNet ResEmoteNet is an innovative and state-of-the-art architecture developed by researchers to improve accuracy and efficiency in this research field. The tool combines CNNs, Squeeze and Excitation networks (SENet) and residual connections to effectively, efficiently and robustly capture the seven basic emotions according to Ekman (1992). The CNN has three hierarchical feature extraction layers, an SE block that models the relationships between the channels and increases representational power of the CNN, several residual blocks that simultaneously enable deeper training and address the well known problem of vanishing gradients [35].

The model was evaluated using the three widely used open source datasets FER2013, RAF-DB and AffectNet. The developed model achieved an accuracy of 79.79% on FER2013, 94.76% on RAF-DB and 72.93% on AffectNet (seven instead of eight emotions), which is an improvement compared to current state-of-the-art methods. This makes ResEmoteNet currently the most accurate and efficient model for emotion classification on the above datasets [35].

The implementation was performed using PyTorch and tested on two different hardware configurations, including a MacBook Pro (M2-Pro chip) and a NVIDIA Tesla P1000 GPU demonstrating the flexibility and scalability of the system for different application scenarios [35].

Multi-Scale Vision Transformer (MViT-CnG) The Multi-Scale Vision Transformer with Contrastive Learning (MViT-CnG) is a novel approach in the field of facial expression-based emotion recognition presented in November 2024. It is especially designed to improve emotion recognition by addressing the challenges posed by different age groups and subtle facial dynamics [6]. Unlike traditional CNNs, MViT-CnG combines vision transformer architecture with multi-scale processing and contrastive learning, enabling superior performance

in capturing both global and fine-grained features in facial expressions.

The term *multi-scale* in MViT refers to the model's ability to analyze input images at multiple resolutions or scales. For instance, it processes low-level details, such as edges and textures, while also capturing high-level semantic information like overall facial expressions. This approach ensures that the model can simultaneously extract local and global features, which is crucial for recognizing nuanced emotional cues across diverse facial expressions. Contrastive learning is used to improve the feature representation of the model. In this technique, the model is trained to minimize the similarity between different emotional expressions (negative pairs), while maximizing the similarity between variations of the same expression (positive pairs). For example, two slightly rotated images of a smiling face are treated as a positive pair, while an image of a smiling face and one showing anger form a negative pair. Furthermore, preprocessing techniques, such as brightness adjustment, flipping and translation help to achieve the model's aim. This allows the model to generalize better, ensuring robustness in recognizing emotions across varying contexts and age groups.

Evaluated on FER-2013 and CK+ datasets, MViT-CnG demonstrated remarkable accuracy: 99.6% on FER-2013 and 99.5% on CK+. Additionally, the high precision, recall and F1-Scores indicate a good real-world applicability and not just an overfitting to a high accuracy. These results show that the approach can cope better with age-related variations and recognize subtle facial expressions, outperforming traditional CNN-based models [6]. This makes it suitable for applications requiring high precision, such as mental health monitoring and adaptive learning environments.

The MViT-CnG represents a significant advance in FER by utilizing multi-scale vision transformers and contrastive learning. It overcomes the limitations of existing CNN-based methods by focusing on both global and local features, making it a robust and versatile tool for diverse real-world applications [6]. However, limitations can still be observed as this model approach has not been tested with datasets other than the FER-2013 and CK+, and the interpretability of the model decisions is less transparent due to the use of transformer-based architectures.

Emotion-LLaMa Emotion-LLaMA is an powerful multi-modal large language model specifically build for emotion recognition and reasoning. It incorporates information from audio, visual, and textual modalities using specific encoders such as HuBERT for audio and multi-view visual encoders like MAE, VideoMAE, and Enhanced Visual Attention (EVA) to provide a comprehensive analysis of emotional cues. Subsequently, the extracted features from audio, video and text are mapped to a shared dimensional space using a linear projection mechanism. Different combinations of encoders were tested, with the best combination being HuBERT, MAE (static facial expression features), VideoMAE

(temporal dynamics) and EVA for global visual context. The model is trained on the MERR dataset created by the authors and enriched with 28,618 coarse-grained and 4,487 fine-grained emotional annotations. This results in coverage of a wide range of emotional categories [13].

Emotion-LLaMA achieves significant benchmarks in emotion recognition tasks. With an F1 score of 0.9036 on MER2023 (Multi-modal Emotion Recognition challenge), it has demonstrated state-of-the-art performance and set new records in zero-shot and fine-tuning evaluations on datasets such as DFEW, achieving the highest Weighted Average Recall (59.37%) and Unweighted Average Recall (45.59%). The results outperform similar approaches like the GPT-4V and Video-LLaMA when compared in emotion recognition and reasoning. These results highlight the effectiveness of the approach in recognizing subtle emotional expressions, including micro-expressions, and vocal nuances, which are critical for real-world applications [13].

Two mentioned limitations are the dataset bias, as certain emotions are still underrepresented, as well as the computational resources needed, especially for the processing of multi-modal data [13]. The model is open-source and available on GitHub for further research and application, enabling accessibility and collaborative improvements in the field. Emotion-LLaMA's combination of multi-modal integration, comprehensive dataset enrichment, and state-of-the-art performance makes it a robust solution for emotion recognition in domains such as education, healthcare, and human-computer interaction.

4.2 Overview of the Tool Comparison Approach

To create a meaningful and comprehensive comparison, five tables are created, focusing on different aspects of the tools.

First, general information is gathered, looking at used modalities and emotion models, indication on the open source availability and the primary use cases. Second, technical aspects such as the model architectures and frameworks are compared. In addition, the input and output data types, usage of pre-trained models and the real-time capability of the analyzed tools are considered. The third table contains information on the reported Key Performance Indicators (KPI's) that are achieved using the respective tool, as well as the datasets on which they are measured. Following, the implementation and ease of use are compared based on the ease of setup, supported programming languages, integration support and the quality of the found documentation. The last table summarizes the findings, highlighting key strengths and limitations and providing a recommendation regarding their usage.

General Information Starting with the General Information (Table 1), the goal is to provide a first overview of the eight analyzed tools. A broad selection of

modalities are utilized: openSMILE processes audio, OpenFace (2.0), DeepFace, LibreFace and the Multi-Scale ViT analyze image and video data, ResEmoteNet focuses on static images. The Emotion-LLaMA integrates multimodal data, including audio, video, and text. The emotion model used criteria indicates either the emotion model integrated into the tool itself, or the model used in research applying the tool for emotion recognition (first four tools). Ekman-FACS is a method for analyzing facial movements based on 26 AU's, while Ekman-Based refers to broader emotion recognition grounded in Ekman's emotion categories. Seven of the eight selected tools are open-source, with the exception of Multi-Scale ViT. However, it is included due to its transformer-based approach and high accuracy on two well-known datasets.

The primary use case of the first four tools is the extraction of features from images/video or audio files to enable the recognition of emotions, while the latter tools are developed for direct classification.

Technical Aspects and Model Architecture The comparison table focusing on technical aspects and model architectures (Table 2) enables a deeper understanding while remaining as comparable as possible. OpenFace (2.0), DeepFace, and ResEmoteNet are all based on CNN's in different forms of design and composition. LibreFace combines transformer-based approaches (Swin Transformer, MAE) with the ResNet-18 CNN. The Multi-Scale Vision Transformer and Emotion-LLaMA both rely on attention-based transformer architectures. In contrast, openSMILE is build using a modular feature extraction pipeline to analyze audio files. Concerning the frameworks, OpenFace (2.0), ResEmoteNet, LibreFace and Emotion-LLaMA make use of the PyTorch package. The Multi-Scale Vision Transformer is created in MATLAB, DeepFace draws upon Tensorflow & Keras, while openSMILE is written in C++. All tools except the Multi-Scale Vision Transformer incorporate pre-trained models, and openSMILE provides predefined feature sets. The input data types align with the used modalities, while the output data types depend on the primary use case. The first four tools focus on extracting feature sets, while the remaining tools provide emotion classification as output. Additionaly, the Emotion-LLaMA is able to provide reasoning for the detected emotion. Lastly, the real-time capability is currently not achieved for ResEmoteNet, Multi Scale Vision Transformer and Emotion-LLaMA, as these tools prioritize accuracy improvements and multimodal approaches.

Performance Metrics The comparison of the performance (Table 3) shows a fundamental comparison problem. The different models use inconsistent KPI's and metrics (accuracy, F1 Scores and correlation coefficients) as well as different data sets, which blurs direct conclusions about the actual accuracy. While DeepFace (FER+: 90.25%) and MultiScale-ViT (FER-2013: 99.6%) appear to achieve high classification accuracies, these values are based on partially non-overlapping data sets, which limits generalizing statements. In addition, the choice of correlation metrics makes it difficult to classify LibreFrace and

OpenFace (2.0). LibreFace provides an average Pearson correlation coefficient (0.63 compared to other models), while OpenFace 2.0 works with the Concordance Correlation Coefficient (0.73 compared to other models), two statistically not directly comparable measures. Table 3 also shows that Emotion-LLaMa was validated multimodally (MER2023 with an F1 score of 0.9036) and that ResEmoteNet achieves an accuracy of 94.76% on RAF-DB, while the accuracy for AffectNet is 79.79% , indicating that the performance of the models is highly dependent on the training data and testing context.

These inconsistencies make it clear that a valid classification of the models would only be possible through a separately conducted study with harmonized testing conditions (same data sets, metrics), meaning that this study was only able to provide a general assessment of the models.

Implementation and Ease of Use Implementationand Ease of Use (Table 4) provides a comparison of the framework or model properties and shows a clear preference for Python as the chosen programming language, while the systems differ significantly in terms of setup effort and integration options in existing systems and pipelines. It quickly becomes clear that the complexity of setting up the system, most of the tools were assessed as easy or fairly easy , but the two newer approaches Multi-Scale ViT and Emotion-LLaMa require more advanced technical knowledge to implement. The use of Python as a programming and interface language, combined with C++ and C# extensions in four tools, reflects the industry wide focus on flexibility and performance. While Liberface is characterized by its cross-platform compatibility, DeepFace and ResEmoteNet focus on the connection to existing Python libraries whilst Multi-Scale ViT does not offer any standardized integration options. The quality of the documentation was categorized as comprehensive for seven out of eight tools, which significantly reduces reproducibility, training time and implementation time for further research. This table illustrates the trade-off between simplicity and innovation. Established frameworks such as DeepFace, LibreFace and Openface rely on pragmatism, while modern models such as Multi-Scale ViT and Emotion-LLaMa require higher implementation costs despite innovative methods.

Summary and Recommendation Table 5 shows a comparison of the most important strengths and limitations as well as the potential application areas of the analyzed FER tools. It shows the dichotomy between open source flexibility and domain specific restrictions, with LibreFace providing a robust basis for real-time applications due to its efficiency and documentation quality despite platform limitations. While LibreFace offers more extensive documentation and cross-platform connectivity, the limitation of the OpenSense component to Windows systems represents a challenge for Linux oriented workflows. The dominance of open source solutions quickly becomes clear and underlines the focus on reproducibility, with established tools such as DeepFace and LibreFace enabling low barrier experiments and studies. Beside LibreFace, DeepFace, OpenFace 1.0

and 2.0 offer real-time capability through pre-trained models and prototype-friendly features, which is a key advantage, especially in an academic context. Modern approaches such as Multi-Scale ViT, which is characterized by superior classification accuracy, and the multimodal-oriented Emotion-LLaMa are not real-time capable, which limits their use to research intensive scenarios.

The table also shows a clear target group separation, meaning that modern solutions such as ResEmoteNet, Multi-Scale ViT and Emotion-LLama focus on intense academic applications with very high accuracy, while openSMILE, OpenFace, DeepFace and LibreFace address the needs of developers for modularity and fast implementation.

4.3 Implementation

In this paper, the open-source tool LibreFace was implemented to demonstrate the potentials of FER. It is optimized for real-time analysis and allows reproducible, cross-platform integration. As a Python based solution, it offers a seamless connection to common data science workflows, which simplifies the development of prototype applications in the research field of information systems. This ensures compatibility with established libraries such as TensorFlow or OpenCV, allowing them to be integrated faster into existing pipelines. The comprehensive documentation and the availability of pre-trained models also reduce the configuration effort, which is crucial for the scientific validation of the results. In Addition, it was also documented how the models can be further or newly trained if desired.

LibreFace can be used as an OpenSense component in a no-code setup. OpenSense is an open source platform for multimodal real-time analysis of social signals and is based on Mirosoft's Platform for Situated Intelligence (psi)[40]. The CPU only version was chosen to ensure reproducibility of this work. The entire pipeline can be imported as a provided JSON file. You can use LibreFace out of the box with just a few adjustments in the settings (detailed instructions in the provided GitHub repository). In addition, OpenSense offers the option of extending or modifying the existing pipeline with other components such as openSMILE. Although a few components of LibreFace are currently only supported on Windows, the advantages of the open source architecture tradeoff the disadvantages, especially the transparency of the models, the efficiency of the emotion classification and the extraction of the feature sets as a .csv file. It generates both qualitative emotion labels and quantitative feature vectors that enable multidimensional evaluation in various research scenarios (Figure 2). LibreFace simplifies research in the FER field and can be implemented as a no-code setup for a variety of studies and can be easily modified for the specific purpose.

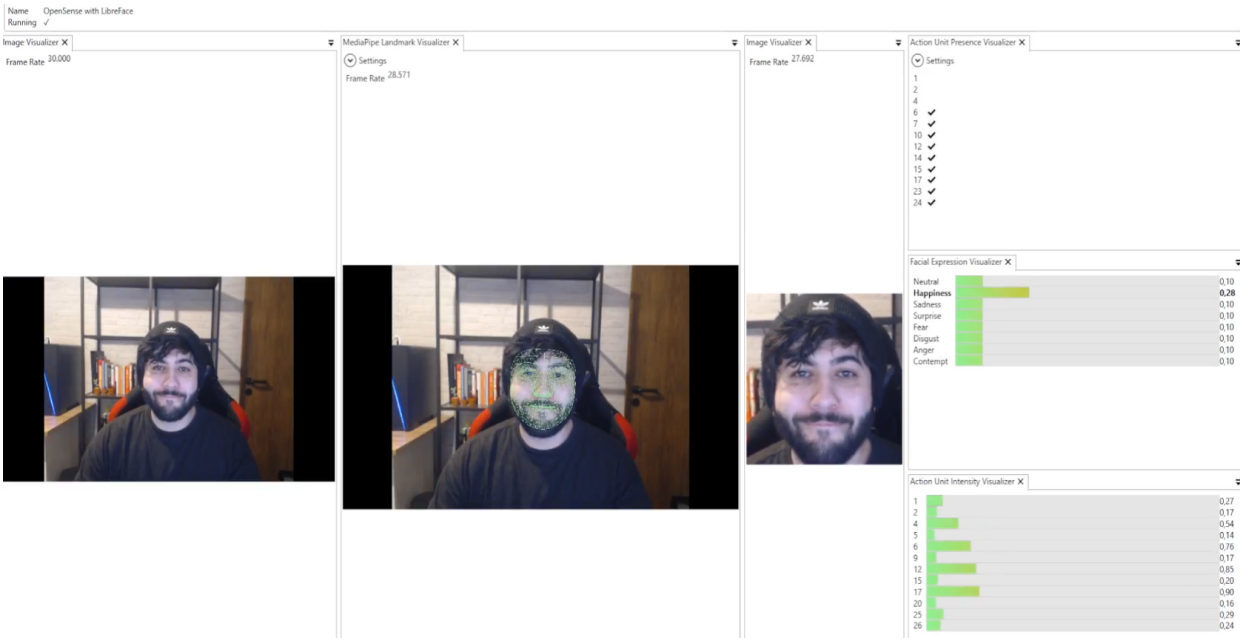

Fig. 2. LibreFace implementation as a OpenSense component

5 Discussion and Critical Reflection

Our approach to evaluating emotion recognition tools highlights both strengths and challenges. One key strength is the broad selection of tools analyzed, covering both single- and multimodal approaches. These include CNN-based models, transformer-based architectures, and hybrid methods. This variety allows for a more comprehensive understanding of how different technologies contribute to emotion recognition. Additionally, the comparison tables provides a structured overview of the tools' strengths, limitations, and real-time capabilities, making it easier to identify the best-suited options for different applications. The practical implementation of LibreFace further demonstrates how emotion recognition can be applied in real-time scenarios, balancing feasibility with technical complexity and performance.

However, several challenges arise in this field. Tool-specific issues include data imbalance and quality, as many datasets lack diversity, especially regarding cultural differences and underrepresented emotional categories. Computational complexity is another concern, as real-time emotion recognition requires significant processing power, making some models difficult to implement efficiently. Furthermore, when it comes to implementation, ethical concerns related to privacy, bias, and user consent are universal across all tools but were not the main

focus of this comparison.

Beyond these technical aspects, there were also personal challenges during this research. The dynamic research landscape presented a major challenge, as the field is evolving rapidly, with new models and improvements constantly emerging. This is particularly evident as three of the eight selected tools were only published in late 2024, making it challenging to assess their long-term reliability and performance.

This leads to another challenge: balancing scientific validity with emerging trends. Many of the latest tools have not undergone extensive validation, yet they demonstrate significant potential. Finding a balance between relying on well-established methods and incorporating cutting-edge approaches required careful evaluation.

Lastly, KPI comparison challenges complicated the analysis. Differences in benchmarking methodologies, a variety of used datasets, and performance reporting made direct comparisons between tools difficult. Despite these challenges, this study provides a structured comparison of emotion recognition tools, highlighting both their potential and areas where further research is needed.

6 Conclusion

This study provides a comprehensive overview of the current state of the art in emotion recognition, focusing on its modalities, technical features and real-time capabilities. The evaluation emphasizes the role of basic emotion models such as Valence-Arousal and Ekman-based models and their integration into widely used tools such as OpenFace, DeepFace and LibreFace.

By comparing eight selected tools, this thesis emphasizes the real-time feasibility and practical implementation. A prototype using LibreFace demonstrates the balance between technical complexity and feasability, showing real-time emotion classification in action.

We suggest that future research should explore new technologies, including Vision Transformers and LLaMA-based models, to improve accuracy and efficiency. In addition, enhancing multimodal approaches can further advance emotion recognition systems and make them more adaptable for real-world applications.

References

1. TechDispatch: facial emotion recognition; issue 1, 2021. https://doi.org/10.2804/014217, last accessed 2025/02/14 (2021)
2. Ahmed, N., Aghbari, Z.A., Girija, S.: A systematic survey on multimodal emotion recognition using learning algorithms. Intelligent Systems with Applications **17**, 200171. https://doi.org/10.1016/j.iswa.2022.200171 (2023)
3. Akçay, M.B., Oğuz, K.: Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication **116**, 56–76. https://doi.org/10.1016/j.specom.2019.12.001 (2020)
4. Amos, B., Ludwiczuk, B., Satyanarayanan, M.: OpenFace: A general-purpose face recognition library with mobile applications (2016)
5. Awana, A., Singh, S.V., Mishra, A., Bhutani, V., Kumar, S.R., Shrivastava, P.: Live emotion detection using deepface. In: 2023 6th International Conference on Contemporary Computing and Informatics (IC3I), pp. 581–584. IEEE. https://doi.org/10.1109/IC3I59117.2023.10397747 (2023)
6. Balachandran, G., Ranjith, S., Chenthil, T.R., Jagan, G.C.: Facial expression-based emotion recognition across diverse age groups: a multi-scale vision transformer with contrastive learning approach 49(1), 11. https://doi.org/10.1007/s10878-024-01241-8 (2024)
7. Baltrusaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607 (2019)
8. Baltrusaitis, T., Robinson, P., Morency, L.P.: OpenFace: An open source facial behavior analysis toolkit. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–10. IEEE. https://doi.org/10.1109/WACV.2016.7477553 (2016)
9. Bruin, J., Stuldreher, I.V., Perone, P., Hogenelst, K., Naber, M., Kamphuis, W., Brouwer, A.M.: Detection of arousal and valence from facial expressions and physiological responses evoked by different types of stressors 5, 1338243. https://doi.org/10.3389/fnrgo.2024.1338243 (2024)
10. Calvo, R.A., D'Mello, S.: Affect detection: An interdisciplinary review of models, methods, and their applications. IEEE Transactions on Affective Computing **1**(1), 18–37. https://doi.org/10.1109/T-AFFC.2010.1 (2010)
11. Chamishka, S., Madhavi, I., Nawaratne, R., Alahakoon, D., De Silva, D., Chilamkurti, N., Nanayakkara, V.: A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling. Multimedia Tools and Applications **81**(24), 35173–35194. https://doi.org/10.1007/s11042-022-13363-4 (2022)
12. Chang, D., Yin, Y., Li, Z., Tran, M., Soleymani, M.: LibreFace: An open-source toolkit for deep facial expression analysis. https://doi.org/10.48550/ARXIV.2308.10713, https://arxiv.org/abs/2308.10713, version Number: 2 (2023)
13. Cheng, Z., Cheng, Z.Q., He, J.Y., Sun, J., Wang, K., Lin, Y., Lian, Z., Peng, X., Hauptmann, A.: Emotion-LLaMA: Multimodal emotion recognition and reasoning with instruction tuning. https://doi.org/10.48550/arXiv.2406.11161, http://arxiv.org/abs/2406.11161 (2024)
14. De la Torre, F., Cohn, J.F.: Facial expression analysis. In: Moeslund, T.B., Hilton, A., Krüger, V., Sigal, L. (eds.) Visual Analysis of Humans, pp. 377–409. Springer London. https://doi.org/10.1007/978-0-85729-997-0_19 (2011)

15. Ekman, P.: An argument for basic emotions 6(3), 169–200. https://doi.org/10.1080/02699939208411068 (1992)
16. Ekman, P., Friesen, W.V.: Facial action coding system. https://doi.org/10.1037/t27734-000, https://doi.apa.org/doi/10.1037/t27734-000, institution: American Psychological Association (1978)
17. Eyben, F., Weninger, F., Gross, F., Schuller, B.: Recent developments in openSMILE, the munich open-source multimedia feature extractor. In: Proceedings of the 21st ACM international conference on Multimedia, pp. 835–838. ACM. https://doi.org/10.1145/2502081.2502224 (2013)
18. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM international conference on Multimedia, pp. 1459–1462. ACM. https://doi.org/10.1145/1873951.1874246 (2010)
19. Girard, J.M., Cohn, J.F., Mahoor, M.H., Mavadati, S., Rosenwald, D.P.: Social risk and depression: Evidence from manual and automatic facial expression analysis. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp. 1–8. IEEE. https://doi.org/10.1109/FG.2013.6553748 (2013)
20. Google: MediaPipe-lösungsleitfaden, https://ai.google.dev/edge/mediapipe/ solutions/guide?hl=de (2024)
21. Gunes, H., Schuller, B.: Categorical and dimensional affect analysis in continuous input: Current trends and future directions. Image and Vision Computing **31**(2), 120–136. https://doi.org/10.1016/j.imavis.2012.06.016 (2013)
22. Hamm, J., Kohler, C.G., Gur, R.C., Verma, R.: Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. Journal of Neuroscience Methods **200**(2), 237–256. https://doi.org/10.1016/j.jneumeth.2011.06.023 (2011)
23. Hossain, M.S., Muhammad, G.: Emotion recognition using secure edge and cloud computing. Future Generation Computer Systems **504**, 589–601. https://doi.org/10.1016/j.ins.2019.07.040 (2019)
24. Hutto, D.D., Robertson, I., Kirchhoff, M.D.: A new, better BET: Rescuing and revising basic emotion theory 9, 1217. https://doi.org/10.3389/fpsyg.2018.01217 (2018)
25. Jack, R.E., Garrod, O.G.B., Yu, H., Caldara, R., Schyns, P.G.: Facial expressions of emotion are not culturally universal 109(19), 7241–7244. https://doi.org/10.1073/pnas.1200155109 (2012)
26. Kaggle, FER-2013 Dataset https://www.kaggle.com/datasets/msambare/ fer2013 (2020)
27. Kaur, J., Saxena, J., Shah, J., Fahad, Yadav, S.P.: Facial emotion recognition. In: 2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), pp. 528–533. IEEE. https://doi.org/10.1109/CISES54857.2022.9844366 (2022)
28. Ko, B.: A brief review of facial emotion recognition based on visual information 18(2), 401. https://doi.org/10.3390/s18020401 (2018)
29. Li, B., Lima, D.: Facial expression recognition via ResNet-50 2, 57–64. https://doi.org/10.1016/j.ijcce.2021.02.002 (2021)
30. Liu, T., Yuan, X.: Paralinguistic and spectral feature extraction for speech emotion classification using machine learning techniques 2023(1), 23. https://doi.org/10.1186/s13636-023-00290-x (2023)

31. Ludwiczuk, B.: History of face recognition: Part 1, https://medium.com/@melgor89/history-of-face-recognition-part-1-deepface-94da32c5355c (2024)
32. McDaniel, B., D'Mello, S., King, B., Chipman, P., Tapp, K., Graesser, A.: Facial features for affective state detection in learning environments (2007)
33. Mehta, D., Siddiqui, M., Javaid, A.: Facial emotion recognition: A survey and real-world user experiences in mixed reality 18(2), 416. https://doi.org/10.3390/s18020416 (2018)
34. Rieber, R.W. (ed.): Encyclopedia of the History of Psychological Theories. Springer US, New York, NY. https://doi.org/10.1007/978-1-4419-0463-8 (2012)
35. Roy, A.K., Kathania, H.K., Sharma, A., Dey, A., Ansari, M.S.A.: ResEmoteNet: Bridging accuracy and loss reduction in facial emotion recognition. https://doi.org/10.48550/ARXIV.2409.10545, https://arxiv.org/abs/2409.10545, version Number: 2 (2024)
36. Russell, J.A.: A circumplex model of affect 39(6), 1161–1178. https://doi.org/10.1037/h0077714 (1980)
37. Sariyanidi, E., Gunes, H., Cavallaro, A.: Automatic analysis of facial affect: A survey of registration, representation, and recognition 37(6), 1113–1133. https://doi.org/10.1109/TPAMI.2014.2366127 (2015)
38. Shaikh, Kazi, Jasani, Sawant, Shaikh, Jinturkar: A survey on facial emotion recognition and fake emotion detection techniques 20(6), 2758–2767. https://doi.org/10.52783/jes.3284 (2024)
39. Shehu, H.A., Sharif, M.H., Uyaver, S.: Facial expression recognition using deep learning. p. 070003. https://doi.org/10.1063/5.0042221 (2021)
40. Stefanov, K., Baiyu, Zongjian, Li, Soleymani, Mohammad: OpenSense: A Platform for Multimodal Data Acquisition and Behavior Perception. https://github.com/ihp-lab/OpenSense (2020)
41. Stratou, G., Scherer, S., Gratch, J., Morency, L.P.: Automatic nonverbal behavior indicators of depression and PTSD: Exploring gender differences. In: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, pp. 147–152. IEEE. https://doi.org/10.1109/ACII.2013.31 (2013)
42. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: DeepFace: Closing the gap to human-level performance in face verification. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1708. IEEE. https://doi.org/10.1109/CVPR.2014.220 (2014)
43. Tracy, J.L., Randles, D.: Four models of basic emotions: A review of ekman and cordaro, izard, levenson, and panksepp and watt 3(4), 397–405. https://doi.org/10.1177/1754073911410747 (2011)
44. Turcian, D., Stoicu-Tivadar, V.: Speech emotion recognition using openSMILE and GPT 3.5 transformer. In: Digital Health and Informatics Innovations for Sustainable Health Care Systems, pp. 924–928. IOS Press. https://doi.org/10.3233/SHTI240562 (2024)
45. Venkatesan, R., Shirly, S., Selvarathi, M., Jebaseeli, T.J.: Human emotion detection using DeepFace and artificial intelligence 59(1), 37. https://doi.org/10.3390/engproc2023059037 (2023)
46. Yuan, X., Wong, W.P., Lam, C.T.: Speech emotion recognition using multi-layer perceptron classifier. In: 2022 IEEE 10th International Conference on Information, Communication and Networks (ICICN), pp. 644–648. https://doi.org/10.1109/ICICN56848.2022.10006474 (2022)

Appendix

Table 1: General Information

Criteria	**openSMILE**	**OpenFace**	**OpenFace 2.0**	**DeepFace**	**ResEmoteNet**	**LibreFace**	**Multi-Scale ViT**	**Emotion-LLaMA**
Primary Modality	Audio	Image/Video	Image/Video	Image/Video	Image	Image/Video	Image/Video	Multimodal(Audio, Video, Text)
Emotion Model Used[1]	Ekman-Based[2]	Ekman-FACS[2]	Ekman-FACS[2]	Ekman-FACS[2]	Ekman-Based[2]	Ekman-FACS[2]	Ekman-Based[2]	Valence-Arousal, Ekman-Based[2]
Open Source?	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
Primary Use Case	Audio Feature Extraction, Paralinguistics	Real-time Facial Feature Analysis	Real-time Facial Feature Analysis	Face Verification & FER	FER	Real-time FER	High-Accuracy FER	Multimodal ER & Reasoning

[1] Indicates either the model integrated into the tool itself, or the model used in research applying the tool for emotion recognition.
[2] Ekman-FACS: Method for analyzing facial movements; Ekman-Based: Broader emotion recognition grounded in Ekman's basic emotion.

Table 2: Technical Aspects & Model Architecture

Criteria	**openSMILE**	**OpenFace**	**OpenFace 2.0**	**DeepFace**	**ResEmoteNet**	**LibreFace**	**Multi-Scale ViT**	**Emotion-LLaMA**
(Model) Architecture	Feature Extraction Pipeline (modular)	CLNF (instance of CLM) + CNN	CLNF (instance of CLM) + CNN	CNN (8 layers) + LCL, 3D face model	CNN + SENet + ResNets	ResNet-18, Swin Transformer, MAE	ViT+ multi-scale processing + contrastive learning	LLaMA + emotion-specific encoders
Framework Used	C++	PyTorch	PyTorch	Tensorflow & Keras	PyTorch	PyTorch	MATLAB	PyTorch
Pre-Trained Models?	No; pre-defined feature sets	Yes	Yes	Yes	Yes	Yes	No	Yes
Input Data Type	Audio files	Image/Video	Image/Video	Image/Video	Image	Image/Video	Image	Video, Audio, Text
Output Data Type	Feature sets	Feature sets	Feature sets	Feature sets	Emotion Classification	Feature sets + Emotion Classification	Emotion Classification	Emotion Classification and Reasoning
Real-Time Capability	Yes	Yes	Yes	Yes	No	Yes	No	No

Table 3: Performance Metrics

Criteria	**openSMILE**	**OpenFace**	**OpenFace 2.0**	**DeepFace**	**ResEmoteNet**	**LibreFace**	**Multi-Scale ViT**	**Emotion-LLaMA**
Reported KPI's (%)	EMO-DB: 83.97% Romanian: 74% IEMOCAP: 60.87%	Average concordance correlation coefficient on DISFA: 0.70	Average concordance correlation coefficient on DISFA: 0.73	FER+: 90.25%	FER-2013: 79.79% RAF-DB: 94.76% AffectNet: 72.93%	Pearson Correlation Coefficient Avg. of AU compared to other methods **LibreFace**: 0.63, **OpenFace2**: 0.59	FER-2013: 99.6% CK+: 99.7%	F1 score: 0.9036 on MER2023-SEMI, UAR: 45.59%, WAR: 59.37% on DFEW
Datasets Used	EMO-DB, IEMOCAP, MELD, Romanian EMO-DB	LFPW, Helen, Multi-PIE	LFPW, Helen	FER+, CK+, LFW, Youtube Faces	FER-2013, RAF-DB, AffectNet	EmotioNet, AffectNet, FFHQ, RAF-DB, DISFA	FER-2013; CK+	MER2023, MER2024, DFEW, EMER

Table 4: Implementation & Ease of Use

Criteria	openSMILE	OpenFace	OpenFace 2.0	DeepFace	ResEmoteNet	LibreFace	Multi-Scale ViT	Emotion-LLaMA
Ease of Setup	Moderate	Fairly Easy	Easy	Fairly Easy	Fairly Easy	Easy	Complex	Complex
Programming Language Support	C++ native, Python integration	C++, C#, Python	C++, C#, Python	Python	Python	C#, Python	MATLAB	Python
Integration Support	ML-Tools (WEKA, SVM, Python libraries)	C++ or C# based project	C++ or C# based project	API compatible with Python libraries	Compatible with Python libraries	Cross-platform support	No	Compatible with Python-based pipelines
Documentation Quality	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Basic	Comprehensive

Table 5: Summary & Recommendation

Criteria	openSMILE	OpenFace	OpenFace 2.0	DeepFace	ResEmoteNet	LibreFace	Multi-Scale ViT	Emotion-LLaMA
Key Strength	Real-Time + batch Support, Open Source, feature library, modular & integrable	Open-source, real-time performance, comprehensive facial behaviour analysis	Open-source, real-time performance, comprehensive facial behavior analysis	Open-source, real-time, pre-trained models	Open-source, efficient, effective	Open-source, efficient, effective, documentation, possible integrations	Superior accuracy	Integration of audio, visual, and text inputs; robust performance across diverse datasets
Key Limitations	Requires expertise for setup, relies on external classifiers for ER	Optimal results at least 100×100 px	Optimal results at least 100×100 px	Computationally heavy due to LCL	Complexity, no real-time support	As OpenSense component only available for Windows	Complexity, no real-time support	Complexity due to multimodality, no real-time support
Recommendation/Best for	Detailed Audio Analysis	Real-Time FA for Researchers and developers	Real-Time FA for Researchers and developers	Real-Time FA for Researchers and developers	*Accuracy* > *Speed* (academic application)	Real-time Video ER	*Accuracy* > *Speed* (academic application)	Research and applications requiring nuanced emotion recognition and reasoning

GitHub Repository

https://github.com/BirulZ/UserStateDetection