Emotion detection through leveraging Action Units and Facial Landmarks while watching video content

Geanovu Medeea-Elena, Gheorghe Octavian-Mihail, Tibrea Mihai-Ionuț

Coordinator: PhD Dr. Kuderna-Iulian Bența

Abstract

Facial Action Unit (AU) recognition is a critical component of emotion detection systems, enabling fine-grained analysis of facial expressions. Traditional methods rely on extensive feature extraction from dense facial landmarks or external datasets. In this paper, we present a lightweight and efficient approach to AU recognition and emotion classification using a selection of 17 facial landmarks. Our program leverages MediaPipe for landmark extraction, OpenFace for generating a labeled training dataset with AUs, and OpenCV for real-time frame capture.

The main focus of the application resides in the implementation of a Convolutional Neural Network (CNN) to detect AUs based on minimal landmark data alongside a Support Vector Machine (SVM) for emotion classification trained on an identical structure of landmarks. Results demonstrate that our method achieves high accuracy while significantly reducing computational complexity, making it suitable for real-time applications.

In our knowledge of the existing study cases, this is among the first efforts to implement both AU detection and emotion detection on such a limited set of landmarks.

Introduction

Understanding facial expressions is a fundamental part of creating a system capable of analyzing human emotions. Knowing how people feel while watching videos is important for creating better content, whether for marketing, education, or entertainment. By recognizing emotions in real-time, creators can see what parts of their videos engage or affect viewers the most. This project aims to develop an app that can detect and analyze users' emotions as they watch videos, and as such, be able to also use the emotional tone to influence the viewers' moods, and consequently, their decisions.

Video content related to specific topics (e.g. charitable actions, product advertisements, personal stories, commentaries) elicits distinct emotional responses that can be captured through real-time analysis. These emotional responses can be used to monitor viewers' mental states which in turn, can help prevent through their report influence on immediate decisions, like mood shifts, or longer-term choices, such emotional state, behavior or opinion changes.

The app uses advanced facial recognition technology to monitor facial expressions and identify emotions like happiness, anger, sadness, surprise, fear,

disgust and contempt. By processing these expressions quickly, the app provides immediate feedback on how viewers are reacting. This real-time analysis helps in understanding viewer engagement in relation to the content they are viewing.
Automatic Facial Expression Analysis (AFEA) has become a prominent study subject. Concerning our proposed idea, by decoding facial cues, it may be possible to develop software able to delve deeper into insights concerning the relationships and interconnectedness of the long periods of time spent browsing social media, the content viewed by the user, and changes caused by the consumption of that said content.

Traditional systems often rely on dense facial landmark data or external feature extraction for Action Unit (AU) recognition and emotion classification. These approaches, while effective and extensively tested to ensure accuracy, can be computationally intensive and require extensive labeled datasets for training that can consume long periods of time. More so, there is limited work on using as little landmark data as possible for lightweight yet robust emotion recognition systems that can be deployed both on desktop and mobile platforms.

In this work, we propose a novel pipeline for real-time AU detection and emotion classification using only 17 facial landmarks. Our system employs a Convolutional Neural Network (CNN) to process landmark data and predict active AUs, a decision tree for evaluating AU combinations to infer emotions, and a Support Vector Classifier (SVC) with Softmax outputs for probabilistic emotion predictions. The results from these channels are combined, accentuating the importance of the CNN's and SVC's outputs over the expected emotion label during the final phase of emotion selection to enhance classification accuracy. By leveraging OpenCV for real-time video capture and MediaPipe for advanced and precise landmark extraction, our pipeline achieves a compact and efficient solution for emotion recognition.

To the best of our knowledge, this is one of the first works to address AU recognition and emotion classification using a minimalistic landmark-based approach. Our contributions include a lightweight architecture, a promising combination of data processing channels for emotion detection, and an evaluation of the system's performance across the training dataset and original input.

Emotion Dataset

For training and testing our emotion detection model, we used the AffectNet dataset, which consists of facial expression data collected from thousands of individuals across various demographic groups. The chosen dataset contains images (~29000) labeled with the range of emotions concerning our project idea, counting happy, angry, sad, fear, surprise, disgust, contempt and neutral. To better fit the dataset for our specific task of emotion detection, we applied several preprocessing steps.

To prepare the training dataset, we ran an Action Unit (AU) detection algorithm using the OpenFace library, specifically employing the Haar Cascade Frontal Face model for face detection. After detecting the faces, we applied the AU detection algorithm, keeping only specific AUs (AU01_r, AU02_r, AU04_r, AU05_r, AU06_r, AU07_r, AU09_r, AU10_r, AU12_r, AU14_r, AU15_r, AU17_r, AU20_r, AU23_r, AU25_r, AU26_r, AU45_r, where r - intensity of the AU), which are commonly associated with emotion rich facial expressions.

Additionally, we performed landmark extraction, but instead of using all facial landmarks, we selected a subset of landmarks: the nose tip, inner and outer corners of the eyes, inner and outer corners of the eyebrows, corners of the mouth, forehead, and chin. These landmarks were normalized with respect to the nose's coordinates to ensure consistent feature representation across various faces and avoid erroneous translations of the landmark coordinates when they will be processed by the CNN model.

For the testing dataset, we recorded real-time facial expressions from the project members using OpenCV and extracted facial landmarks using MediaPipe's holistic model. The holistic model, with a minimum detection confidence of 0.5 and a minimum tracking confidence of 0.5, was employed to extract the x and y coordinates, data corresponding to each frame was saved every 3 seconds, allowing for analysis of emotions during certain time frames and comparing the result of the emotion classification with the expected label of that timestamp.

The video dataset consisted of short videos available on social media platforms, across different markets and demographics (entertainment, ads, news). These videos were used to compare the output of the emotion classification pipeline and the emotion associated with the content presented in the videos.

Method

In this section, we present the proposed methodology for emotion classification using facial landmarks and Action Units (AUs). The model processes a frame depicting a participant's face and outputs an emotion label based on the detected facial expressions. The analysis is divided into three stages: preprocessing, AU detection with the aid of the CNN, emotion probability detection by using the AU correlated with emotions decision tree and the SVC, and emotion classification.

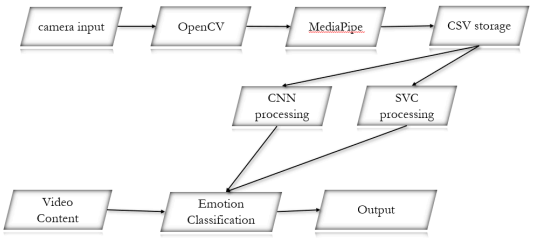


Fig. 1 shows an overview of the architecture.

1. Preprocessing

The input to the model consists of a set of landmarks corresponding to facial features (e.g. nose, eyes, mouth, eyebrows, cheeks).

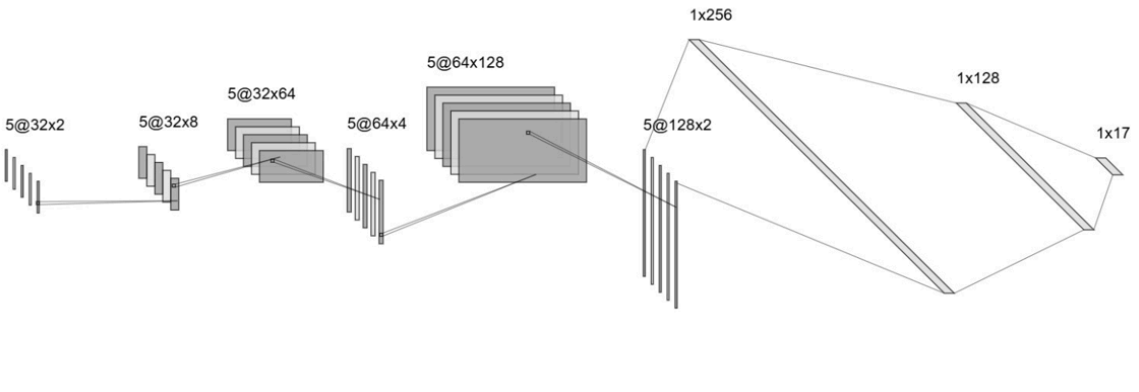
Face Detection and Alignment: We use OpenCV and MediaPipe's holistic model to detect and extract the participant's face from each frame. The images are rescaled and the channel is changed to RGB. This model also provides a set of 3D landmarks, but for the current implementation, only the x and y coordinates are extracted. To reduce variation caused by the positioning of the participant's face, we align the face to a neutral position using the nose as a reference.

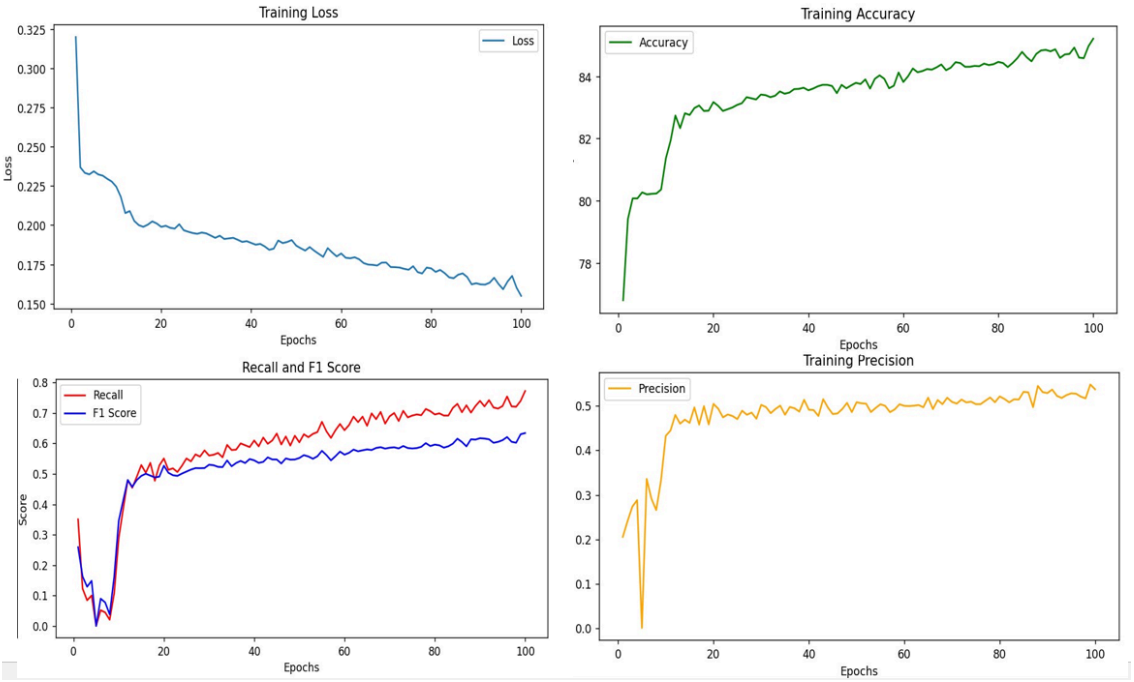
Normalization: To ensure consistency across different participants and frames, the landmark coordinates are normalized with respect to the nose's coordinates, scaling them within a fixed range of [-1,1]. Any missing entries from the dataset, infinite or null values are changed with 0.

$$\tilde{x}_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Fig. normalization function applied on the landmark coordinates, where xi is the current coordinate and x is the landmark array.

Fig. and Fig. CNN structure and views of the model's performance.





Landmark Selection: We focus on a subset of 17 facial landmarks (e.g. nose tip, eye corners, mouth corners, etc.), which are more relevant for detecting AUs in relation to the limited number of landmarks we imposed as the available data, and minimizing computational complexity.

2. AU Detection

In this stage, we use a Convolutional Neural Network (CNN) to predict the presence or absence of AUs. It receives the extracted vector of landmarks from the opencv/mediapipe image processor. These vectors, as above mentioned, have x, y and the name of the coordinate. The CNN is able to detect the action units from those coordinates, as it was trained on the AU annotated dataset previously mentioned at chapter Emotion Dataset.

The CNN architecture consists of several convolutional layers followed by a fully connected layer that outputs a vector

of probabilities for the presence of 17 AUs. The model is trained to output values in the range [0,1], after which values over 0.5 are considered valid activations of the AU, where 1 indicates the activation of a specific AU and 0 indicates its absence.

AU Dataset and Training: For training, we used the AffectNet dataset, where each frame is labeled with the presence or absence of various AUs. We used the subset of AUs that are most relevant to emotion recognition by analysing the chosen landmarks. The CNN was trained using the landmark vectors from the dataset and the corresponding AU labels.

AU Detection Model: The CNN architecture is a 1D-CNN with multiple layers that extracts spatial relationships and maps them to AUs.
Input Layer: Receives normalized facial landmark data in the shape (batch_size, num_landmarks, 2), where 2 corresponds to (x, y) coordinates.

Convolutional Layers: Extract local spatial patterns between consecutive landmarks, complexity increases the deeper in the network we are.

Conv1: Input: (batch_size, 2, num_landmarks) Output: (batch_size, 32, num_landmarks/2) (after pooling) Detect local patterns between consecutive landmarks (eg, eyebrow movement due to the eye corner and outer eyebrow landmarks being neighbours).

Conv2: Input: (batch_size, 32, num_landmarks / 2) (after pooling) Output: (batch_size, 64, num_landmarks / 4) Purpose: Learn more complex relationships by combining local patterns. Kernel size will be 3.

Conv3: Input: (batch_size, 64, num_landmarks / 4) (after pooling) Output: (batch_size, 128, num_landmarks / 8) Purpose: Learn more complex relationships by combining local patterns. Kernel size will be 3.

Pooling Layers: MaxPool1d: Reduces the size of the feature map by half, retaining the most important features. Helps the model focus on key patterns and reduces computational complexity.

Fully Connected (Dense) Layers:
FC1: Input: Flattened feature map from the convolutional layers. Output: 128 neurons, each representing a high-level feature. Activation: ReLU introduces non-linearity.

FC2: Further processes the 256 features and outputs 128 intermediate features. Activation: ReLU.
FC3 (Output Layer): Maps the 128 features to the number of AUs (e.g. 17 in this case). Activation: Sigmoid, ensuring each output neuron provides a probability score between 0 and 1.

Activation Functions: ReLU (Rectified Linear Unit): Used in hidden layers to introduce non-linearity and prevent vanishing gradients. Sigmoid: Used in the output layer for multi-label classification, providing a probability score for each AU, threshold being 0.5, so the end result is a binary vector with 0 for not present and 1 for certainly present.

Training: For training, we use the Binary Cross-Entropy Loss function to compute the loss of the model as it is suitable for multi-label classification. The optimizer of the presented model is the Adam Optimizer for its adaptive learning rate.

3. Emotion Classification

In this stage, we use two methods to classify the detected emotions based on the AUs and landmark coordinates.

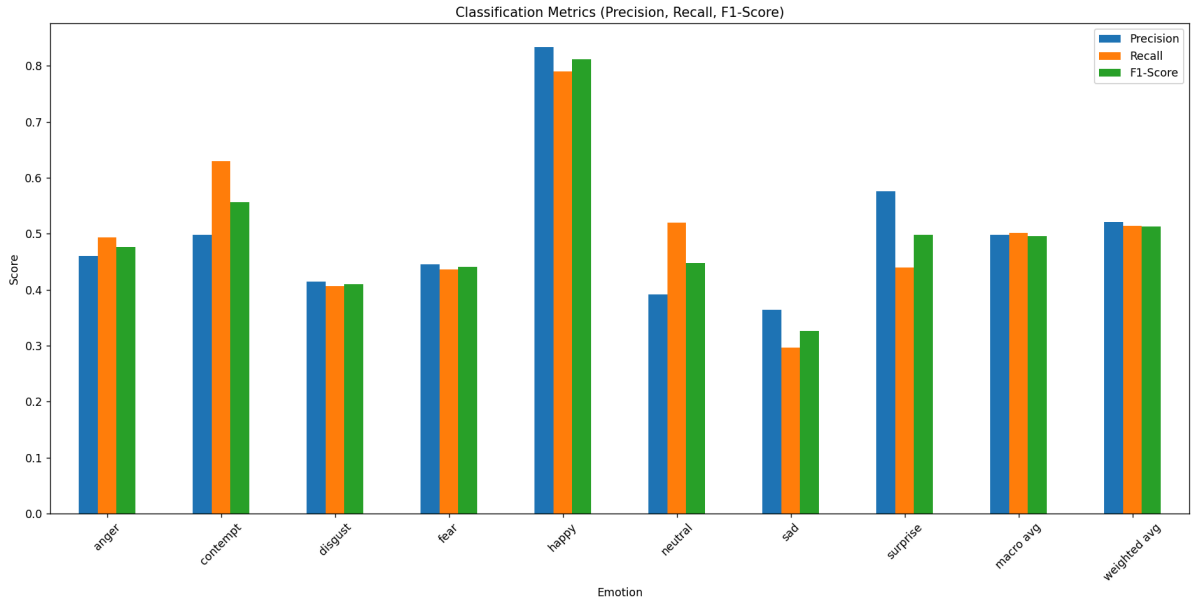


Fig. The performance of the SVC model.

3.1 Decision Tree

After obtaining the AU predictions from the CNN, we use a decision tree to classify the emotions based on the combination of active AUs. The decision tree evaluates different combinations of active AUs and outputs an emotion label. This method helps in inferring the most likely emotion, with happiness, anger and sadness being the best recognized emotions.

3.2 Support Vector Classifier (SVC)

The Support Vector Classification receives the dataset consisting of annotated images with their landmarks (the x,y coordinates). The landmarks are flattened to a 1D array and all their respective labels are appended to another array. The flattened landmarks are then standardized. We run the SVC training loop with options to learn automatically its best parameters. Afterwards, we use the most accurate SVC model to directly classify emotions from the raw landmark coordinates. The SVC takes the normalized landmark vector as input and outputs a probability distribution over the possible emotions. We apply Softmax to the SVC's output to obtain a probability distribution that sums to 1, where each value corresponds to the likelihood of a specific emotion.

3.3 Combining Results

To combine the outputs from the CNN (AU predictions) and the SVC (emotion probabilities), we first use the decision tree to process the AU predictions obtained from the output of the CNN, to derive an emotion label based on the active AUs. The SVC provides a probability distribution of emotions directly from the landmark coordinates. The final emotion label is determined by evaluating the combined results of both the CNN and SVC, considering 0.4 of the CNN's emotion probability and 0.4 of the emotion probabilities of the SVC output.

4. Evaluating the AU and SVC models

The performance of the CNN and SVC models was evaluated using the following set of metrics, including

accuracy, precision, recall, F1-score, and confusion matrix analysis.

$$Recall = \frac{TP}{TP + FN} \quad Precision = \frac{TP}{TP + FP}$$

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$

Fig. Formulas of the metrics used to measure the models' validity and generalization.

CNN Results				
Loss: 0.1549,	Accuracy: 0.85%	Precision: 0.54	Recall: 0.77	F1: 0.63
SVC Results				
C: 100, Kernel: auto	Accuracy:: 0.51%	Precision: 0.50	Recall: 0.51	F1: 0.51

Fig. The overall performance of the used classification methods.

Conclusions

This project proved to have average levels of accuracy, while the requirement of efficiency in computation time was met. Using a small number of significant facial landmarks to detect the emotion of facial expressions through AUs combinations and SVC classification has promising potential. The application's generated insights can be valuable for detecting the effects of video content in marketing, social media, or emotional manipulation situations and understanding audience reactions, which will aid in forming a report on the psychological effects of exposure to the aforementioned cases on the audience with further development.

References

[1] Affectiva, "Website: https://www.affectiva.com/," Techniques: Face Detection and Tracking, Emotion Detection, Feature Extraction, Deep Learning, Algorithms: CNNs trained on large datasets, Facial Feature Extraction, Approach: Emotion AI embedded in applications for audience reaction analysis.

[2] Realeyes, "Website: https://www.realeyesit.com/," Techniques: Computer Vision, Face Detection, Attention Measurement, Algorithms: CNNs for emotional valence and arousal detection, clustering for pattern identification, Approach: Supervised learning for emotion recognition, unsupervised methods for emotional trends.

[3] Microsoft Azure Face API, "Website: https://azure.microsoft.com/en-us/services/cognitive-services/face/," Techniques: Cloud Computing, Multi-task Learning, Algorithms: CNNs for facial recognition and emotion classification using transfer learning, Approach: Classifies emotions such as happiness, sadness, anger, and surprise.

[4] Emotient, "Website: https://www.researchgate.net/figure/Analyzing-different-emotions-on-face-using-Emotient-Tool-interviews_fig1_312964373/," Techniques: FACS, Psychological Integration for Emotion Detection, Algorithms: SVMs for Action Unit detection, CNNs for nuanced facial expressions, Approach: Psychology-driven machine learning models for robust emotion recognition.

[5] EmotionSense, "Website: https://www.emotionsense.org/," Techniques: Multimodal Emotion Recognition, Audio-Based Sentiment Analysis, Algorithms: Deep neural networks for Feature Fusion and Emotion Classification, Approach: Adapts to user-specific behavior with supervised and unsupervised learning.

[6] nViso, "Website: https://www.nviso.ai/," Techniques: 3D Facial Modelling, Feature Extraction, Algorithms: CNNs for facial landmark detection, Bayesian Networks for uncertainty management, Approach: Real-time emotion recognition with robust accuracy under varying conditions.

[7] Mustafa Okan Irfanoglu, Berk Gokberk, Lale Akarun, "3D Shape-based Face Recognition Using Automatically Registered Facial Surfaces," Link: https://www.researchgate.net/publication/4090704_3D_Shape-based_face_recognition_using_automatically_registered_facial_surfaces

[8] Mina Bishay, Ahmed Ghoneim, Mohamed Ashraf, Mohammad Mavadati, "Which CNNs and Training Settings to Choose for Action Unit Detection? A Study Based on a Large-Scale Dataset," Link: https://arxiv.org/pdf/2111.08320

[9] Philipp Michel, Rana El Kaliouby, "Facial Expression Recognition Using Support Vector Machines," Link: https://www.cs.cmu.edu/~pmichel/publications/Michel-FacExpRecSVMPoster.pdf

[10] Essam Haider Mageed, Hind Rustum Mohammed, Asaad Norri Hashim, "Novel System for Face Recognition Based on SVD and GLCM," Link: https://www.researchgate.net/publication/318640470_Novel_System_for_Face_Recognition_Based_on_SVD_and_GLCM

[11] Bruce Poon, Ashraful Amin, Hong Yan, "PCA-Based Human Face Recognition with Gradientfaces," Link: https://www.iaeng.org/IJCS/issues_v43/issue_3/IJCS_43_3_02.pdf

[12] Narendra Kumar Rao B; Nagendra Panini Challa; E S Phalguna Krishna; S. Sreenivasa Chakravarthi, Facial Landmarks Detection System with OpenCV Mediapipe and Python using Optical Flow (Active) Approach," Link: https://ieeexplore.ieee.org/document/10182585/

[13] Maninderjit Singh; Anima Majumder; Laxmidhar Behera, "Facial Expressions Recognition system using Bayesian Inference," Link: https://ieeexplore.ieee.org/document/6889754

[14] Seena Jose, Prof. Shivapanchashari, "Unsupervised Method for Face Photo-Sketch Synthesis and Recognition," Link: https://www.ijariit.com/manuscripts/v3i3/V3I3-1177.pdf

[15] Darshan Gera, Naveen Siva Kumar Badveeti, Bobbili Veerendra Raj Kumar, S Balasubramanian, "Dynamic adaptive threshold based learning for noisy annotations robust facial expression recognition," Link: https://arxiv.org/pdf/2208.10221

[16] Mina Bishay, Jay Turcot, Graham Page, Mohammad Mavadati, "Automatic Detection of Sentimentality from Facial Expressions," Link: https://arxiv.org/pdf/2209.04908

[17] Chih-Jen Lee; Tzu-Yin Chen; Jenn-Dong Sun; Tai-Ning Yang; Allen Y. Chang, "Combining Gradientfaces, principal component analysis, and Fisher linear discriminant for face recognition," Link: https://ieeexplore.ieee.org/document/5573238

[18] Wei Sun, Anshul Sheopuri, ing Li, Thales Teixeira, "Computational Creative Advertisements ," Link: https://dl.acm.org/doi/pdf/10.1145/3184558.3191549

[19] Taiping Zhang, Yuan Yan Tang, Bin Fang, "Gradientfaces for Illumination-Invariant Recognition," Link: https://www.researchgate.net/publication/220502996_Face_Recognition_Under_Varying_Illumination_Using_Gradientfaces