JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

INTELLIGENT REAL-TIME FACIAL EMOTION DETECTION USING COMPUTER VISION AND DEEP LEARNING

A fused landmark, CNN and YOLOv8-based system with a live telemetry dashboard

[1]Renuka V Deolalikar, [2]Prof Vijaykumar M P
[1]Student, [2]Assistant Professor
[1]Computer Science and Engineering
[1]Shreeyash College of Engineering and Technology (SYCET), Chhatrapati Sambhajinagar, India

Abstract : Real-time facial expression analysis supports human–computer interaction, e-learning and driver-monitoring research, but systems that rely on one classifier are unstable and report only seven coarse emotions. This paper presents a real-time, web-based system that estimates a basic emotion (seven classes) and a fine-grained emotion (the 27 GoEmotions categories plus neutral) directly from a webcam stream. The browser extracts 468 face landmarks with MediaPipe Face Mesh and streams them, with a small face crop, to a FastAPI back end over a WebSocket. The back end fuses a landmark-geometry rule set calibrated to each person's resting face with an image classifier trained on FER2013 (a custom CNN, with ResNet18, MobileNetV3 and YOLOv8-cls options) and smooths the result over time. The fine emotion is chosen inside the recognised basic group by matching seven facial cues (smile, mouth opening, eye widening, brow position, lip narrowing, head pitch and head tilt) to cue prototypes of the fine labels. A dashboard also reports blinks, gaze, head pose, an attention-engagement score and throughput stability. An audit of FER2013 found 703 training images that also occur in the test set, and the training pipeline removes them. A GoEmotions text model reached 64.2% top-1 agreement with the labelled basic group on held-out comments, and the earlier custom CNN reached 61.95% accuracy on the FER2013 test set. The fine-emotion mapping is rule-based and has not been validated on fine-labelled faces.

IndexTerms **- Facial emotion recognition, fine-grained emotion, MediaPipe Face Mesh, FER2013, GoEmotions, convolutional neural network, real-time system, attention estimation.**

I. INTRODUCTION

Much of the emotional information exchanged in everyday interaction is carried by the face, and automatic recognition of facial expressions has applications in human–computer interaction, online learning, healthcare support and driver monitoring [1], [2]. Most systems classify a face into the basic emotions proposed by Ekman [1] and are trained on datasets such as FER2013 [3]. Two limitations remain. First, a single classifier applied frame by frame is unstable: its predictions flicker between neighbouring classes and are sensitive to lighting, head pose and the gap between benchmark images and webcam video. Second, seven labels are coarse; people also describe states such as amusement, nervousness, remorse or disappointment, which the text-based GoEmotions taxonomy captures with 27 categories plus neutral [4].

No openly available face dataset carries such fine-grained labels, so a fine-grained facial classifier cannot simply be trained on the data used for basic emotions. This work therefore uses the two available resources in different ways: the image dataset trains the basic-emotion classifiers, and the fine-grained taxonomy is the label space for a cue-based mapping that selects a fine label inside the recognised basic group.

This paper extends our earlier real-time system, which combined YOLOv8 face localisation, MediaPipe Face Mesh and a CNN for seven-class emotion recognition with blink, gaze, head-orientation and attention analysis [5]. The present version localises the face from the MediaPipe landmarks and uses YOLOv8 only as an optional emotion classifier. The contributions of this paper are: (i) a real-time architecture in which landmark extraction runs in the browser and emotion inference runs in a Python service connected by a WebSocket; (ii) a probabilistic fusion of landmark rules and image classifiers with temporal smoothing and a per-person resting-face baseline; (iii) a facial-cue method that selects one of the GoEmotions labels directly from the face, without user input; (iv) a telemetry module for head pose, blinks, gaze, attention and throughput that is covered by automated tests; and (v) an audit of FER2013 that quantifies duplicated and conflicting images and a leak-free training split.

II. Related Work

Facial expression research is rooted in the Facial Action Coding System (FACS), which describes expressions as combinations of muscle movements called action units [6]. Basic-emotion theory [1] underlies most benchmark datasets, including FER2013, a collection of 35,887 grayscale images of 48×48 pixels labelled with seven emotions [3]. Convolutional networks, and more recently networks pre-trained on ImageNet such as ResNet [7] and MobileNetV3 [8], are the standard way to classify such images. FER2013 has noisy labels, and human accuracy on it is only about 65% [3].

Landmark-based approaches use geometric measurements of facial points instead of pixels. MediaPipe Face Mesh predicts 468 three-dimensional landmarks in real time on commodity hardware [9], [10] and, with the refined option, also provides iris landmarks. Landmark ratios such as the eye aspect ratio are widely used to detect blinks [11], and the percentage of eye closure (PERCLOS) is an established drowsiness indicator [12]. Dimensional models of affect, such as the valence–arousal circumplex [13], motivate describing expressions with graded cues rather than only with discrete labels.

For fine-grained emotion, GoEmotions provides about 58,000 Reddit comments annotated with 27 emotion categories plus neutral [4]. It is a text resource without paired face images, so fine-grained labels cannot be learned directly from faces with it. The psychological literature also cautions that facial movements do not map reliably to inner emotional states across people and contexts [2]; the outputs of the system described here are therefore presented as estimates of visible expression.

III. System Design and Methodology

System overview

Figure 1 shows the architecture. The browser captures webcam video and runs MediaPipe Face Mesh [9] (one face, refined landmarks, detection and tracking confidence 0.5). It computes head pose, blinks, gaze, an attention score and throughput locally, and sends a message to the back end at most every 150 ms (about 6–7 messages per second) containing the 468 landmarks, the head pose relative to the person's baseline and, every 400 ms, a 112×112 JPEG crop of the face. The back end is a FastAPI service run by Uvicorn; each WebSocket connection owns an emotion engine, so the resting-face baseline, the smoothing state and the fine-label state belong to one person. The engine returns the basic emotion, the fine emotion with its runner-up, the top basic-emotion probabilities, a confidence value and the source of the decision, and the dashboard renders them live. A CSV log of the session is written once per second.

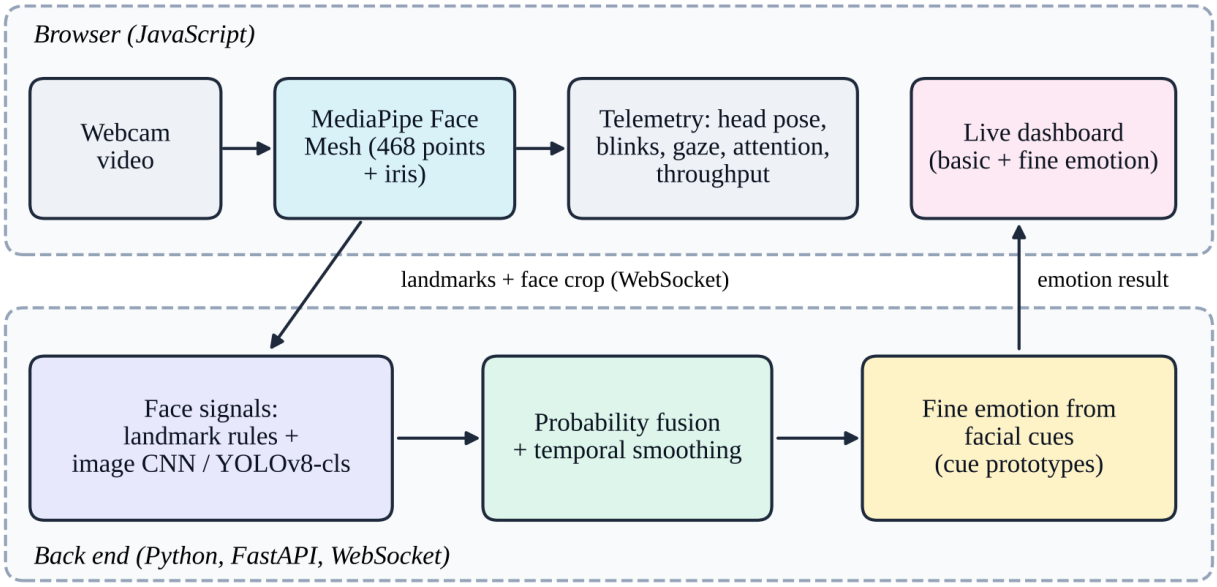

Fig. 1. System architecture of the proposed real-time emotion detection system

Datasets

Two datasets are used (Table 1). FER2013 contains 28,709 training and 7,178 test images of 48×48 grayscale faces in seven folders: angry, disgust, fear, happy, neutral, sad and surprise [3]. The classes are imbalanced: disgust has 436 training images (1.5%), 16.5 times fewer than happy (7,215). GoEmotions contains 211,225 annotation rows (one per rater) for 58,011 distinct comments and 28 labels [4]. Rows flagged as very unclear (3,411) were dropped, the annotations were aggregated per comment, and a label was kept when at least 40% of the comment's raters selected it; this left 54,502 labelled comments.

Table 1. Datasets used in this work

Dataset	Content	Size used	Role in the system
FER2013	48×48 grayscale faces, 7 basic emotions	28,709 train / 7,178 test (26,912 train after de-duplication)	Trains and evaluates the image classifiers
GoEmotions	Reddit comments, 27 emotions + neutral	54,502 labelled comments	Fine-label taxonomy; trains an optional text model

Landmark cues and resting-face baseline

Five geometric features are computed from the landmarks and divided by the face width (the distance between landmarks 234 and 454), so that they do not depend on the distance from the camera: mouth opening, mouth width, mouth-corner lift, eye opening and brow-to-eye distance. For each feature k a slow average s_k and a fast average f_k are maintained (smoothing factors 0.015 and 0.35

per message). The deviation of a feature from the person's resting face is its fast average minus its slow average. The slow average is updated only while all deviations are below 0.010, so that a held expression is not absorbed into the baseline; if a deviation persists for about 45 s, the baseline is allowed to drift slowly so that the system recovers from a baseline captured during an expression. Each basic emotion receives a score built from one dominant cue and one or two supporting cues using piecewise-linear ramps; for example, happiness grows from 0 to 3.0 as the corner lift rises from 0.014 to 0.032 and gains up to 1.2 more from mouth widening. Neutral scores 1.0 plus up to 2.0 when all deviations are small. A non-neutral class must exceed neutral by 1.5 points, otherwise its score is damped by a factor of 0.3, and a softmax with temperature 0.9 converts the scores into probabilities.

Image classifiers and training

The custom CNN takes a 48×48 grayscale face and has five 3×3 convolution layers (32, 32, 64, 64 and 128 filters) with batch normalisation and ReLU, three max-pooling stages and a 256-unit fully connected layer with dropout (0.4 and 0.3), about 1.3 million parameters in total. Two ImageNet-pretrained networks, ResNet18 [7] and MobileNetV3-small [8], can be fine-tuned instead (64×64 input replicated to three channels), and an optional YOLOv8n-cls classifier [14] can be trained on the same folders. At inference, the probabilities of the face crop and of its horizontal mirror are averaged.

Training uses AdamW [15] with weight decay 0.01, a cosine learning-rate schedule (initial rate 1×10^{-3} for the custom CNN and 5×10^{-4} for pretrained networks), label smoothing of 0.1 [16], a class-weighted cross-entropy loss with weights proportional to the inverse square root of class frequency, random flips, ±10° rotations, 6% translations, 0.92–1.08 scaling and brightness/contrast jitter, optional contrast-limited adaptive histogram equalisation (CLAHE) [17], early stopping with a patience of six epochs, and a stratified 10% validation split that selects the best epoch. The test set is evaluated once, after training.

FER2013 contains exact duplicate images. Files were compared by hash: 703 training images also occur in the test set, 73 are duplicates that carry conflicting labels, and 1,021 further copies are repeated within the training set. These are removed before training, which leaves 26,912 of the 28,709 training images and prevents the test score from being measured partly on images seen in training.

Fusion and temporal smoothing

Each available signal (landmark rules, image CNN, optional YOLOv8-cls) supplies a probability for every basic emotion. The fused probability of each emotion is the weighted average of these probabilities, with weight 1.0 for the landmark rules and 1.6 for each image model. An image prediction is reused for 1.5 s, because crops arrive about 2.5 times per second whereas landmarks arrive about 7 times per second. The fused probabilities are smoothed with an exponential moving average (factor 0.30), and the displayed emotion changes only when a new leader exceeds the current emotion's probability by 0.06.

Fine emotion from facial cues

The 27 GoEmotions labels are grouped under the seven basic emotions (for example joy, amusement and excitement under happiness; grief, remorse and disappointment under sadness). Given the recognised basic emotion, the system selects one label of that group from seven facial cues, each normalised to [−1, 1] (to [0, 1] for tilt): smile (mouth-corner lift ÷ 0.032), mouth opening (÷ 0.036), eye widening (÷ 0.020), brow raise (÷ 0.024), lip narrowing (÷ 0.040), head pitch down (÷ 15°) and head tilt (÷ 15°). Every fine label has a prototype of target cue values informed by FACS-based descriptions of expressions [6] (excerpt in Table 2). The label whose prototype has the smallest weighted squared distance from the current cues is selected (cue weights are 0.6 for head pitch and tilt and 1.0 for the others). The relative closeness of the two best prototypes is shown as the runner-up, and a new fine label replaces the shown one only after it wins three consecutive updates. Because no fine-labelled face data are available, the prototypes are hand-set and can be tuned; broad labels such as joy, surprise or anger are the more reliable ones, whereas subtle labels such as caring, pride or realization are weak guesses from a face.

Table 2. Excerpt of the cue prototypes of the fine labels

Basic group	Fine label	Characteristic facial cues
Happiness	Joy	broad smile, slight eye squint
Happiness	Amusement	broad smile, mouth open
Happiness	Excitement	broad smile, wide eyes, raised brows
Sadness	Sadness	frown, slightly narrowed eyes
Sadness	Remorse	frown, head pitched down
Sadness	Grief	deep frown, mouth slightly open
Fear	Fear	wide eyes, raised brows, mouth open
Fear	Nervousness	mildly wide eyes, pressed lips
Anger	Anger	lowered brows, pressed lips, frown
Anger	Annoyance	mildly lowered brows
Surprise	Surprise	mouth wide open, wide eyes, raised brows
Surprise	Confusion	lowered brows, head tilt
Disgust	Disgust	narrowed mouth, lowered brows, frown

Head pose, blinks, gaze, attention and throughput

Head pose is estimated from the face-plane normal, obtained as the cross product of the cheek-to-cheek vector (landmarks 234 to 454) and the forehead-to-chin vector (10 to 152) in pixel units. Pitch and yaw are the vertical and horizontal angles of this normal relative to the camera axis. The median of the first 20 frames defines the person's neutral pose; UP or DOWN is reported beyond ±12° and released below 8° (hysteresis), and the baseline drifts slowly (time constant 20 s) only while the head is clearly centred, so that a sustained look up is not absorbed into it.

The eye aspect ratio [11] is computed in pixel space and averaged over both eyes. An eye is treated as closed when the ratio falls below 72% of the person's own open-eye level and as open again above 85%; a closure of 30–700 ms counts as a blink and longer

closures count as eyes closed. Gaze is the mean position of the two irises between the eye corners; values above 0.62 or below 0.38 are labelled left or right.

The attention-engagement score combines head orientation, gaze, eye closure and blink behaviour. A base value is formed from gaze (weight 0.50), alertness (0.35) and blink behaviour (0.15); it is then reduced by up to 70% for head orientation and by up to 80% for prolonged eye closure, and expressed as a percentage. Head orientation is penalised for yaw beyond 15° and pitch beyond 12°; gaze for an iris offset beyond 0.10 from the centre; and eye closure for a continuous closure longer than 0.4 s (full penalty at 1.6 s). Alertness also falls with the share of the last 30 s spent with the eyes closed (PERCLOS) [12], and the blink term is 1, or 0.5 when the blink rate is below 5 or above 35 per minute after 20 s of observation. The score is smoothed with time constants of 1.0 s (rising) and 0.5 s (falling) and labelled FOCUSED (≥ 75), PARTIAL (≥ 50), DISTRACTED (≥ 25) or AWAY; DROWSY marks high eye closure.

Throughput is measured on unique camera frames only. The frame rate is the inverse of the mean interval over the last 60 frames, and stability falls with the ratio of the standard deviation to the mean of the intervals and with the share of dropped frames; it is STABLE at ≥ 85%, FAIR at ≥ 65% and UNSTABLE below that. The round-trip time to the back end is measured with sequence numbers that the server echoes.

Implementation

The front end uses JavaScript, an HTML5 canvas and MediaPipe Face Mesh. The back end uses Python, FastAPI, Uvicorn, OpenCV and NumPy; the image models use PyTorch and torchvision, and the optional detector uses Ultralytics YOLOv8 [14]. The text model is trained with scikit-learn and run with NumPy only. Docker is used for packaging, and the measurement logic is tested with Node.js.

IV. RESULTS AND DISCUSSION

Text model on GoEmotions

A TF-IDF representation (unigrams and bigrams, 50,000 features, sublinear term frequency) with one-vs-rest logistic regression (C = 4, balanced class weights) was trained on 90% of the labelled comments (49,052) and evaluated on the remaining 5,450. The highest-scoring label was one of the comment's labelled emotions for 54.95% of the comments over the 28 labels, the highest-scoring basic group matched a labelled group for 64.24%, and the macro-averaged F1 score at a 0.5 threshold was 0.449. The deployed model was then refit on all comments. These figures are not comparable with published GoEmotions benchmarks because the label aggregation and the split differ, and the model is a bag-of-words classifier that cannot read sarcasm or context.

Image classifiers

The earlier custom CNN reached 61.95% accuracy on the FER2013 test set. This is in the range expected from a small network on noisy FER2013 labels, but it was measured with the best checkpoint selected on the test set and with the 703 overlapping images present, so it is likely to be slightly optimistic. The v5 networks, trained with the leak-free split, validation-based model selection and transfer learning, are reported in Table 3.

Software verification

The measurement logic was verified with automated tests: 18 unit checks of head pose (synthetic faces rotated by known pitch and yaw, including a face at half size), pose hysteresis, blink counting, attention and throughput; a smoke test that runs the page with simulated camera frames; and nine consistency tests in which synthetic landmark faces with known cue values pass through the engine and return the expected basic and fine labels. These tests show that the processing is wired correctly; they are not accuracy measurements on real faces. A held smile of about 22 s of simulated messages stays detected, whereas the earlier baseline update let a held expression fade into the resting face.

Table 3. Summary of evaluation results

Component	Evaluation	Result
Custom CNN (v4)	FER2013 test, 7,178 images	61.95% accuracy
Text model (held out)	5,450 comments, 28 labels	top-1 in labelled set 54.95%; basic group 64.24%; macro-F1 0.449
Telemetry logic	18 automated unit checks	18 / 18 passed
Engine consistency (synthetic faces)	9 cases	9 / 9 passed

Discussion and limitations

Fusing independent signals and smoothing them over time removes much of the frame-to-frame flicker of a single classifier, and the per-person baseline lets the geometry rules adapt to different faces. The main limitation is that the fine emotion is a rule-based mapping rather than a learned model: no face dataset with fine-grained labels is bundled, so the mapping has not been validated against human judgements. Evaluating it requires fine-labelled faces or ratings from human annotators. A second limitation is the domain gap between FER2013 (48×48 grayscale web images) and webcam crops, so live accuracy is expected to be lower than the benchmark score. Finally, expressions are not reliable evidence of inner feelings [2]; the system estimates visible expression and should not be used for decisions about people. Frames and face crops are processed by the local server, and the session log contains only labels and telemetry values, not images.

V. CONCLUSION AND FUTURE WORK

A real-time system was built that estimates a basic and a fine-grained emotion directly from a webcam by fusing landmark geometry, an image classifier and temporal smoothing, and that reports head pose, blinks, gaze, attention and throughput on a live dashboard. A data audit exposed duplicated and conflicting FER2013 images, which the training pipeline now removes. Future work includes measuring the transfer-learning and YOLOv8-cls models on the leak-free split, training a fine-grained face model on a

dataset with richer labels or valence–arousal annotations, validating the fine-emotion prototypes with human raters, supporting several faces, and adding per-user calibration.

REFERENCES

[1] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3–4, pp. 169–200, 1992.

[2] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychological Science in the Public Interest*, vol. 20, no. 1, pp. 1–68, 2019.

[3] I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *Neural Information Processing (ICONIP 2013)*, Lecture Notes in Computer Science, vol. 8228. Berlin, Germany: Springer, 2013, pp. 117–124, doi: 10.1007/978-3-642-42051-1_16.

[4] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A dataset of fine-grained emotions," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2020, pp. 4040–4054.

[5] R. V. Deolalikar and R. P. Kharjule, "Intelligent real-time facial emotion detection using computer vision and deep learning," *J. Emerging Technologies and Innovative Research (JETIR)*, vol. 13, no. 8, pp. c245–c247, Aug. 2026, paper ID JETIR2608227.

[6] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto, CA, USA: Consulting Psychologists Press, 1978.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[8] C. Lugaresi *et al.*, "MediaPipe: A framework for building perception pipelines," arXiv:1906.08172, 2019.

[9] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann, "Real-time facial surface geometry from monocular video on mobile GPUs," in *CVPR Workshop on Computer Vision for AR/VR*, Long Beach, CA, USA, Jun. 2019, arXiv:1907.06724.

[10] Federal Highway Administration, Office of Motor Carriers, "PERCLOS: A valid psychophysiological measure of alertness as assessed by psychomotor vigilance," Tech Brief FHWA-MCRT-98-006, Washington, DC, USA, Oct. 1998.

[11] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO (version 8.0.0)," software, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[12] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2019.

[13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2818–2826.

[14] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*. San Diego, CA, USA: Academic Press, 1994, pp. 474–485.