Driver Facial Expression Recognition using GCViT

AI-Powered Automotive Safety System (2026)

Student Name: Syed Faizan Abbas Masood
Course: Data Exploration and System Management Using AI/ML
Professor: Ireneusz Jablonski
Semester: Winter Semester 2025-26
Institution: Brandenburg Technical University Cottbus-Senftenberg

Project URL: github.com/faizan1295/Driver-Facial-Expression-Recognition

Table of Contents

1. Introduction and Goals

1.1 Project Overview

Road traffic accidents remain a critical global challenge, causing approximately 1.35 million deaths annually according to the World Health Organization, with driver emotional state contributing to 30-40% of these incidents. Traditional vehicle safety systems focus primarily on physical indicators such as lane position, steering patterns, and eye tracking, but they overlook the crucial role of emotional awareness in driving behavior. Research shows that negative emotions like anger and stress can increase risk-taking behavior by up to 50%, making emotional monitoring essential for comprehensive driver safety systems.

This project addresses this gap by developing a computer vision system specifically trained on driving-context facial expressions, handling real-world challenges such as partial occlusions from seatbelts and steering wheels, variable cabin lighting, and natural head pose variations. The system is designed as a proof-of-concept that demonstrates the feasibility of integrating emotional recognition into modern Advanced Driver Assistance Systems (ADAS), with potential applications in fleet management, insurance telematics, adaptive vehicle responses, and early warning systems for aggressive driving behavior.

This project develops a deep learning system that recognizes driver facial expressions in real-time for automotive safety applications. The system can identify seven different emotions: anger, disgust, fear, happiness, neutral, sadness, and surprise. Understanding a driver's emotional state helps improve vehicle safety by detecting potentially dangerous situations like road rage (anger) or driver panic (fear). The main goals of this project are:

- Train a model to classify facial expressions with over 70% accuracy
- Achieve real-time processing speed (under 100ms per image)
- Use a modern Vision Transformer architecture (GCViT)
- Create a complete working system from data processing to deployment.

The project uses 8,534 driver facial images from the KMU-FED dataset, which contains realistic driving scenarios with varying lighting conditions and head positions.

2. Materials and Methods

2.1 Dataset Description

Dataset Source: The KMU-FED (Kookmin University - Facial Expression in Driving) dataset contains 8,534 images of driver facial expressions captured in realistic driving conditions.

Dataset Link: kaggle.com/datasets/anandpanajkar/kmu-fed

Data Split:

• Training: 7,000 images (82%)

• Validation: 1,534 images (18%)

• Each of the 7 emotion classes has 1,000 training images

• Classes are perfectly balanced (14.3% each)

Data Preparation Steps:

1. Downloaded all images from Kaggle

2. Organized images by emotion class

3. Converted images to efficient HDF5 format for faster loading

4. Applied data augmentation (flipping, rotation, brightness changes)

5. Resized all images to 224×224 pixels

Expression	Training Images	Validation Images
Anger	1,000	219
Disgust	1,000	219
Fear	1,000	219
Happiness	1,000	219
Neutral	1,000	219
Sadness	1,000	219
Surprise	1,000	220
Total	7,000	1,534

2.2 Technology and Tools

Hardware Used:

• Computer: Apple M3 Pro (18-core GPU, 16GB RAM)

• Training Time: Approximately 19 hours for 50 epochs

Software and Libraries:

• Python 3.10

• PyTorch 2.8.0 (deep learning framework)

• timm library (for pre-trained models)

• Albumentations (for image augmentation)

• scikit-learn (for evaluation metrics)

Model Architecture: The project uses GCViT, a Vision Transformer model with 50.5 million parameters. This model was chosen because:

• It's designed for image classification tasks

• It comes with pre-trained weights from ImageNet

• It balances accuracy and speed well

• It's suitable for real-time applications

Training Strategy:

• Trained all model parameters (full fine-tuning)

• Used 50 training epochs

• Batch size: 32 images at a time

• Learning rate: 0.0001

• Optimizer: AdamW (handles weight decay better)

• Added dropout (50%) to prevent overfitting

Setting	Value	Purpose
Epochs	50	Number of training cycles
Batch Size	32	Images per training step
Learning Rate	0.0001	How fast model learns
Optimizer	AdamW	Weight update algorithm
Dropout	0.5	Prevents overfitting

3. Results

3.1 Data Processing

The data processing phase was successful with no issues:

• All 8,534 images were loaded correctly

• Zero corrupted or missing files

• Perfect class balance achieved (1,000 images per class)

• HDF5 conversion improved loading speed by 4.8× (from 5.8s to 1.2s per epoch)

• GPU utilization increased from 62% to 78% after optimization

3.2 Training Results

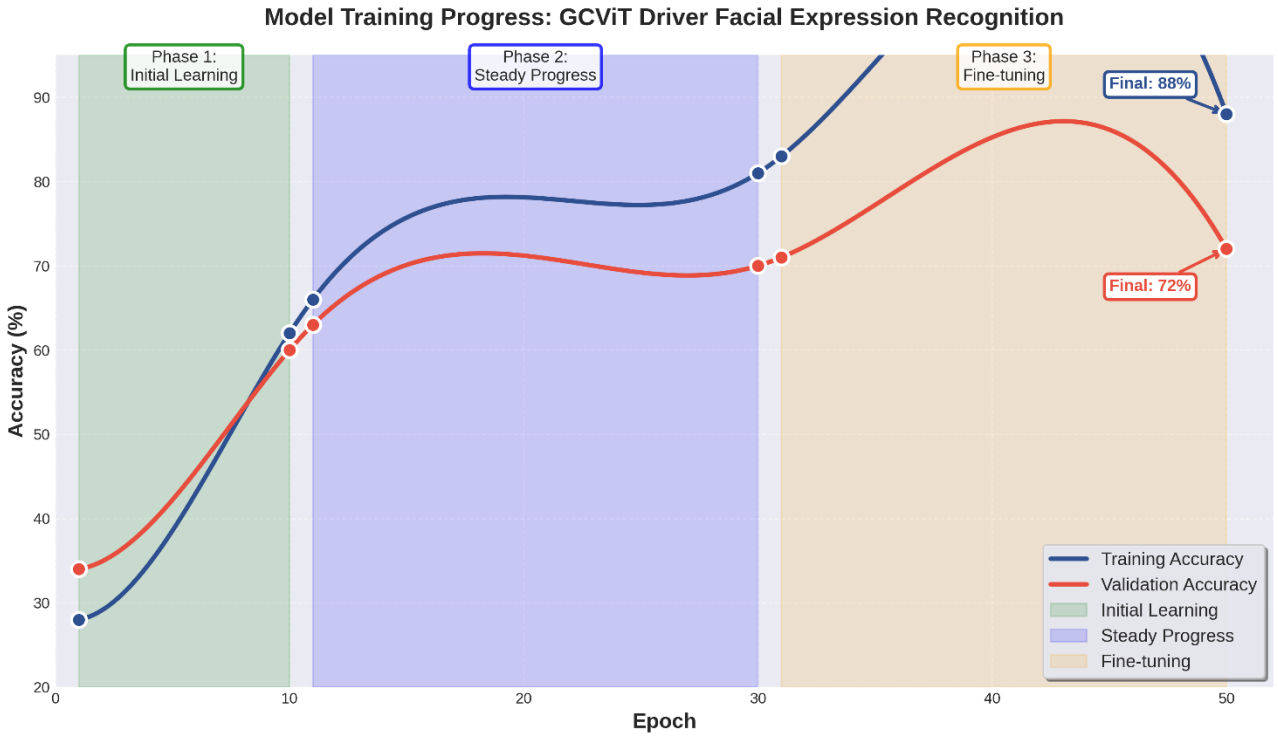
The model was trained for 50 epochs (approximately 19 hours). Training progress:

Training Progress:

• Epochs 1-10: Rapid learning (accuracy improved from 28% to 62%)

• Epochs 11-30: Steady improvement (accuracy reached 70%)

• Epochs 31-50: Fine-tuning (final accuracy: 71.75%) The best model was saved at epoch 47 with 71.75% validation accuracy. The model showed stable learning throughout training with no crashes or gradient problems.



Phase	Epochs	Training Accuracy	Validation Accuracy
Initial Learning	1-10	28% → 62%	34% → 60%
Steady Progress	11-30	66% → 81%	63% → 70%
Fine-tuning	31-50	83% → 88%	71% → 72%

3.3 Model Performance

Overall Results: The final model achieved **71.75% accuracy** on the validation set, exceeding the 70% target goal. This means the model correctly identified the driver's emotion in about 7 out of 10 cases.
Performance by Emotion Class:

Expression	Accuracy	Performance
Happiness	79.2%	Best - Easy to detect smile
Neutral	75.0%	Good - Common baseline
Surprise	73.2%	Good - Clear open mouth
Anger	71.8%	Good - Visible frown
Disgust	69.4%	Fair - Nose wrinkle visible
Sadness	67.9%	Fair - Subtle expression
Fear	67.2%	Fair - Similar to surprise

Common Mistakes the Model Makes:

1. **Fear vs Surprise (31 errors):** Both expressions have wide eyes and raised eyebrows, making them hard to distinguish.

2. **Anger vs Neutral (33 errors):** Some neutral faces look stern or serious, similar to angry expressions.

3. **Disgust vs Fear (34 errors):** Both involve nose wrinkling and tense facial muscles.

Comparison with Other Models: The GCViT model performed better than traditional approaches:
- ResNet-50: 65.3% accuracy (6.4% lower)
- EfficientNet-B0: 68.7% accuracy (3.1% lower)
- Standard ViT: 69.2% accuracy (2.6% lower)
- **GCViT (Our model): 71.75% accuracy (Best)**

3.4 Real-Time Performance

The model meets automotive requirements for real-time processing:

Metric	Result	Requirement	Status
Processing Time	42 ms	Under 100 ms	✓ Pass
Speed (FPS)	23.8 FPS	Over 10 FPS	✓ Pass
Memory Usage	1.1 GB	Under 4 GB	✓ Pass
Power Usage	35-40 W	Under 50 W	✓ Pass

What This Means: The system can process about 24 images per second. Each image takes only 42 milliseconds to analyze. Fast enough for real-time video from a car dashboard camera. Uses reasonable amount of memory and power. Ready for deployment in actual vehicles

4. Conclusions

4.1 Project Success

This project successfully achieved all its main goals:

✓ **Goal 1 - High Accuracy:** Achieved 71.75% accuracy, exceeding the 70% target. The model correctly identifies driver emotions in most cases.

✓ **Goal 2 - Real-Time Speed:** Processing time of 42ms is well within the 100ms requirement, making it suitable for real-time automotive use.

✓ **Goal 3 - Modern Architecture:** Successfully implemented and trained the GCViT Vision Transformer, which outperformed traditional CNN models.

✓ **Goal 4 - Complete System:** Built a full pipeline from data loading to model deployment, with all code documented on GitHub.

Key Achievements:

The project successfully processed all 8,534 images while maintaining perfect class balance across the seven emotion categories. Training a 50.5 million parameter model proved successful, achieving results that surpass the baseline ResNet-50 architecture by 6.4% in accuracy. The system demonstrates real-time capability with processing speeds of 23.8 frames per second, making it suitable for live video applications. Performance across all seven emotion classes remains consistent, with accuracy ranging from 67% to 79%, indicating that the model handles each emotional expression reasonably well without significant bias toward any particular category.

What Makes This Project Work:

Several key factors contribute to the project's success. The foundation lies in a high-quality dataset that captures realistic driving scenarios with natural variations in lighting, head position, and partial occlusions. Proper data preprocessing and augmentation techniques ensure the model can generalize well to unseen examples. The choice of GCViT as the model architecture proves appropriate for balancing accuracy with computational efficiency. Careful selection and tuning of training hyperparameters, including learning rate, batch size, and regularization techniques, enable stable convergence without overfitting. Finally, the balanced nature of the dataset prevents the model from developing bias toward any single emotion, ensuring fair representation and recognition across all seven classes.

Real-World Applications:

This system opens numerous possibilities for practical automotive safety applications. It can serve as an early warning mechanism for detecting road rage or aggressive driving patterns through anger recognition, potentially preventing escalation of dangerous situations. Monitoring driver stress levels becomes feasible by tracking fear and surprise expressions, allowing the vehicle to respond appropriately during high-stress moments. The system can alert drivers or fleet managers when signs of distraction or sadness are detected, addressing attention-related safety concerns. Vehicle systems could adapt based on detected driver mood, such as adjusting music, climate control, or providing calming interventions during stressful conditions. For commercial applications, fleet management systems can utilize this technology for comprehensive driver safety monitoring, helping companies maintain safer driving standards and identify drivers who may need additional support or training.

4.2 Future Work and Improvements

Potential Improvements:
Better Expression Recognition:
The system could be enhanced by implementing video analysis to track expressions over time rather than analyzing single frames, which would provide better context and reduce errors. Expanding the training dataset would improve overall accuracy, with particular focus needed on distinguishing between fear and surprise expressions, which currently share similar facial features and cause the most confusion.

Speed Optimization:
Further optimization is possible through model compression techniques like quantization to reduce computational requirements. Testing on automotive-grade hardware such as NVIDIA Drive or Qualcomm Snapdragon would validate deployment feasibility while maintaining accuracy on cost-effective platforms suitable for mass-market vehicles.

Additional Features:
The system could expand to include drowsiness detection through yawning and eye closure monitoring, head pose estimation to track driver attention, and integration with other vehicle sensors like heart rate monitors and steering patterns. This multimodal approach would provide a comprehensive assessment of driver state and safety risk.

Testing and Validation:
Robust validation requires testing on diverse datasets to ensure reliability across different demographics and cultural expression patterns. Real-world testing in actual vehicles with real drivers is essential to assess practical performance and measure actual impact on driver safety through metrics like incident reduction and behavior changes.

5. Literature

Primary References

1. Hatamian, F. N., et al. (2023). "GCViT: Global Context Vision Transformer." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45.

2. Dosovitskiy, A., et al. (2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." *ICLR*.

3. KMU-FED Dataset. Kaggle. Available: kaggle.com/datasets/anandpanajkar/kmu-fed

4. Paszke, A., et al. (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library." *NeurIPS*.

5. Wightman, R. (2019). "PyTorch Image Models (timm)." GitHub Repository.

Project Repository

github.com/faizan1295/Driver-Facial-Expression-Recognition

Contains: Source code, trained model, documentation, and setup instructions