

Chapter 5: Conclusion and Future Work

1.1. Important Features

All features used within this facial emotion recognition architecture have been carefully designed and optimized to be academic valid, efficient and viable in AT. A system that optimizes the functional aspects of real-time expression mapping while operating on the tight computational budget of an edge deployment, with emphasis on execution time, low latency, and high predictability. Core infrastructure values accessibility and reliable performance in unconstrained, real world environments.

This work is based on the fundamental backbone of Real-Time Emotion Classification using an advanced, customized Sequential Convolutional Neural Network (CNN). We do not rely on the common structures, instead we use our own method of Sequential layout with a customized two-stage pipeline (using data augmentation and automated hard-sample extraction for fine-tuning). The pipeline leverages the Haar Cascade Classifier from OpenCV to automatically detect the face of a specific subject in each incoming frame, crop the region of interest (ROI) to a consistent 48×48 pixels size, and then feed the frame into a neural inference. Expressions are divided into different emotional categories by the CNN: happy, sad, angry, fear, surprise, disgust, and neutral. In assistive scenarios this sub-second execution becomes very important; it provides real-time behavioural insights into the social interaction, enabling people to interpret facial signals correctly and respond readily.

The architecture can potentially be deployed on embedded platforms in future work, as well as web-based platforms powered by the Gradio framework. All these steps, from image acquisition, Haar-cascade preprocessing, CNN inference and audio synthesis, are performed locally, following the edge computing principles. The edge-centric design reduces dependency on constant cloud connectivity, thereby significantly reducing response latency, and naturally ensuring data privacy and security. As a result the application can be used in private and public environments without external servers receiving sensitive biometric data.

In addition, the system includes an automatic auditory feedback loop that converts the predicted emotions into synthesized speech using gTTS (Google

Text-to-Speech) engine. The module conveys a visual expression in a few words and a descriptive phrase, for instance, "Happy" as a classification becomes an expressive spoken phrase: "I am feeling very happy today!"). The system provides context-aware and comprehensible audio feedback, helps users to sense the changes of emotions in their surroundings, decreases communication difficulties, and promotes autonomy in interpersonal communication.

High robustness to environmental variations, including changing illumination, head poses and occluded facial parts. Our deep CNN blocks and strict local preprocessing allows this resilience to be obtained. Furthermore, by optimizing the density of the layers, and adding strategic Dropout regularizations we were able to achieve the highest possible classification accuracy, with the lowest possible computational complexity. The careful balancing makes sure that the emotion recognition and playback of the gTTS happen within an interactive time window, hence making the prototype technically sound and scalable.

1.2. Future Work

The present proposal shows good performance in the emotion classification task at the local level and the responsive voice feedback, but there are several directions that can be followed to enlarge the scope, robustness and analysis depth of the system.

One potential path forward for future work is the transition from custom Sequential CNN architectures to hybrid deep learning approaches, including Vision Transformers (ViTs) and attention-based networks. CNNs are good at extracting local spatial features, and the attention mechanism is suitable for extracting global contextual relationships between facial landmarks. This transition would greatly improve the system's capacity to distinguish between those expressions that are very subtle or ambiguous, such as low intensity fear and surprise or slight sadness and baseline. Under sub-optimal lighting or motion blur.

Furthermore, the use of lightweight pre-trained architectures, designed for edge devices like MobileNetV3 or EfficientNet-Lite, is a promising approach. Transfer Learning techniques from the large scale heterogeneous image data sets can offer the model with richer foundation feature representations. Through this, training time would be cut down while at the same time enhancing the network to generalize its results to a range of people, different ethnicities, age groups and arbitrary environmental settings without the need for massive, localized datasets.

The system can also be expanded from a single-modality vision tracking system to a full-fledged multimodal emotion recognition network. Further future versions would combine facial image analysis with concurrent, parallel behavioral markers such as vocal prosody, speech intonation, and linguistic sentiment monitoring. The fusion of several data streams would greatly enhance the reliability of facial expression classifications, especially when facial expressions are intentionally hidden, ambiguous or partially concealed.

From the hardware/interface architecture perspective, pushing the trained models onto dedicated edge AI accelerators, such as NVIDIA Jetson Nano and Google Coral TPU, would provide additional hardware acceleration. Such migration would allow larger, multi-layered networks to be executed with even less processing latency. At the same time, the user experience could be improved by adding adaptive functionalities to the Gradio UI, such as enabling the user to adapt the language, speed, detail of feedback and tone of the gTTS output to their preference.

Lastly, long-term and structured field studies with target users in dynamic social settings must be done. Field testing outside of a sandbox will reveal real-world issues like adaptation to the system, edge-cases, and reliability over long periods of use. The system will be further developed from an optimized functional prototype to a full-fledged intelligent assistive technology based on these real-world insights, balanced with strict protocols for user privacy and informed data consent.

References

World Health Organization. (2019). *World report on vision*. World Health Organization Press.

Andrada-Livia, C., Popescu, D., & Iordache, D. (2023). New trends in emotion recognition using image analysis by neural networks: A systematic review. *Sensors*, 23(16), 7092. <https://doi.org/10.3390/s23167092>

Hawton, A., Green, C., Dickens, A. P., Richards, S. H., Taylor, R. S., Edwards, R., Greaves, C. J., & Campbell, J. L. (2011). The impact of social isolation on the health status and health-related quality of life of older people. *Quality of Life Research*, 20(1), 57–67. <https://doi.org/10.1007/s11136-010-9717-2>

Hussain, A., Khan, B., & Turkey, A. (2023). A lightweight facial expression recognition system using partial transfer learning for visually impaired people. *IEEE Access*, 11, 1–9. <https://doi.org/10.1109/ACCESS.2023.3234101>

Jung, S.-W. (2019). Facial action units-based image retrieval for facial expression recognition. *IEEE Transactions on Human-Machine Systems*, 49(6), 612–621.

Kirkvik, E. B. (2022). *Facial emotion recognition using deep learning* [Master's thesis, Oslo Metropolitan University]. OsloMet Open Repository.

Kusuma, H., Attamimi, M., & Fahrudin, H. (2020). Deep learning-based facial expressions recognition system for assisting visually impaired persons. *Bulletin of Electrical Engineering and Informatics*, 9(2), 1–12. <https://doi.org/10.11591/eei.v9i2.2081>

Lee, H.-C., Wu, C.-Y., & Lin, T.-M. (2013). Facial expression recognition using image processing techniques and neural networks. *National Quemoy University Journal*, 5(1), 45–58.

Mellouk, W., & Handouzi, W. (2020). Facial emotion recognition using deep learning: Review and insights. *Procedia Computer Science*, 170, 416–423. <https://doi.org/10.1016/j.procs.2020.03.080>

Mitra, A., Choudhury, S., & Moitra, S. (2018). *Image processing facial expression recognition* [Technical Report]. RCC Institute of Information Technology.

Onyema, E. M., Shukla, P. K., Dalal, S., Mathur, M. N., Zakariah, M., & Tiwari, B. (2021). Enhancement of patient facial recognition through deep learning algorithm: ConvNet. *Journal of Healthcare Engineering*, 2021, Article 5514217. <https://doi.org/10.1155/2021/5514217>

Patil, D. D. (2023). *Facial emotion recognition using deep learning* [Master's thesis, California State University, Northridge]. CSUN University Library.

Sari, M., Moussaoui, A., & Hadid, A. (2019). Automated facial expression recognition using deep learning techniques: An overview. *International Journal of Informatics and Applied Mathematics*, 15(2), 1–15.

Shehada, D., Turkey, A., Khan, W., Khan, B., & Hussain, A. (2023). A lightweight FER system using partial transfer learning for visually impaired people. *IEEE Access*, 11, 10214–10225. <https://doi.org/10.1109/ACCESS.2023.3241029>

Sikri, M., Gupta, A., Puri, R., Goel, S., Pathak, N., & Tiwari, M. (2018). Emotion detection using image processing in Python. In *2018 5th International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 5812–5815). IEEE.