Privacy-Preserving Facial Emotion Classification: A Comprehensive Investigation of Federated Learning Approaches

Chun Chen

Department of Applied Physics, Shanghai University of Electric Power, Shanghai, China

SIREN0627@mail.shiep.edu.cn

Abstract. Facial Emotion Recognition and Classification (FERC) is an important technology used in many areas, but it also brings up serious privacy worries because it depends on sensitive facial information. Federated Learning (FL) is a useful way to handle these issues because it lets different systems train a model together without actually sharing their private data. This paper provides a comprehensive review of the integration of FL with FERC. It surveys key FL methodologies--including horizontal, personalized, and cross-silo FL--applied to emotion classification using architectures such as CNNs and ResNet. The review also identifies critical challenges, such as data heterogeneity, the privacy-performance trade-off, and system heterogeneity, which impact model accuracy and practicality. Future research could focus more on personalized and semi-supervised federated learning, explore cross-modal learning, and look into using blockchain to make the systems more trustworthy. This study emphasizes how FL can help keep user data private while still working well in FERC systems. It's a useful example for others looking to improve these kinds of systems in the future.

Keywords: Face Emotion Classification, Federated Learning, Deep Learning, Privacy Protection.

1. Introduction

Facial emotion classification is a technology which automatically recognizes and classifies human emotions by analyzing facial expressions in images or videos. Facial emotion is one of the most immediate and universal forms of non-verbal communication. Happiness, sadness, anger, fear, surprise, disgust are the six basic emotions. The study of facial emotion has important applications in many fields. In the field of medicine, this technology can monitor patients to observe whether they exhibit symptoms of depression [1], anxiety or other emotional disorders [2, 3]. It can prevent traffic accidents by detecting the driver's fatigue and distracted state [4]. Identifying suspicious or abnormal behaviors in crowded areas is also one of the methods of application [5]. However, since human faces are regarded as personal privacy information, it is a serious issue that such data is used for improper purposes. Privacy protection and data security are among the greatest values of Federated Learning (FL). Therefore, it is advisable to consider integrating Facial Emotion Recognition and Classification (FERC) with FL for research and development.

FL has witnessed remarkable progress since its inception, addressing privacy concerns in data-driven AI. Jakub et al. pioneered the FL concept [6], proposing the Federated Averaging (FedAvg) algorithm that aggregates locally trained model updates instead of centralizing raw data. This breakthrough cut communication costs by 10 to 100 times compared to regular SGD, while still working well on tricky datasets like MNIST and CIFAR-10 that have uneven data. It's a big step forward for private distributed learning.

Subsequently, Sawsan et al. looked at the main issues in FL, like how systems are built, keeping data private and managing resources [7]. They also mapped out how FL is used across different fields. Their work gives a clear guide for how to better adopt FL and improve its technology. These advances have enabled FL's integration into sensitive fields like healthcare and finance, reconciling regulatory compliance (e.g. GDPR) with model performance.

There have been some important progressions in using federated learning for facial emotion classification. These new developments help protect privacy better and also make the models work more effectively. Shome et al. introduced FedAffect [8], a few-shot federated learning setup designed for facial expression recognition. It helps models perform better on new, real-world data and allows different clients to work together without sharing sensitive information, which is an important step for real-world use. Researchers developed a decentralized facial emotion classification model integrating FL with CNN, ResNet-50, and VGG-16 architectures, tested on FER-2013 and CK+ datasets. Their CNN model achieved a maximum accuracy of 77%, resolving privacy concerns in decentralized emotion recognition.

Su et al's FedLC framework can work in different setups, but it shows especially strong benefits when it comes to facial emotion classification [9]. It helps improve model transmission in networks where data can get lost, boosting accuracy by nearly 4%, or reducing the amount of traffic needed by about 34-48%. This makes a big difference for real-time FERC systems running on the edge. These works validate FL's role in reconciling privacy needs and performance demands of facial emotion classification.

The aim of this paper is to conduct a summary and review of the previous achievements in this field. In Section 2, the development history of the application of FL on facial emotion classification was described and introduced. In Section 3, the current shortcomings in this field and the future development direction will be presented. Finally, Section 4 is the summary of the entire paper.

2. Method

2.1. Horizontal Federated Learning (HFL)

HFL is applicable when different clients (e.g. smartphones) have the same feature space but different sample IDs for their data. In emotion classification, this means each device has its own collection of facial images annotated with emotion labels. Without sharing the original image data, it can collaboratively train a more powerful and versatile convolutional neural network model.

McMahan et al. showed that a federated-averaged model can achieve accuracy comparable to a model trained on centralized data [10]. For facial emotion classification, this means the final model performs well across different demographic groups or lighting conditions represented in the separate datasets, while preserving patient confidentiality at each hospital.

2.2. Personalized Federated Learning

In emotion recognition, due to differences in culture, demographics or individual expression styles, the data of different users may be completely different. A single global model trained by the standard federated averaging algorithm often performs poorly for individual users.

Arivazhagan et al. [11] manages the base layers, which learn general features (like edges and shapes). The personalization layers remain on the user's device and are trained solely on their local data. For emotion classification, the base layers learn generic facial features, while the personalization layers adapt to the specific emotional expressions and lighting conditions unique to that user. This personalization significantly boosts accuracy compared to a one-size-fits-all global model, making FL practical for real-world applications where user data is inherently diverse.

Fallah et al. proposed Per-FedAvg [12], an algorithm designed to generate a global model that can be quickly personalized to new users with minimal data. Their experiments used fully-connected DNNs on non-vision datasets, but the framework works with any kind of model. Applying Per-FedAvg to facial emotion classification helps a deep neural network quickly develop a strong starting point for recognizing different emotions. When deployed to a new user's device, this model could be fine-tuned efficiently on a small set of the user's personal images, achieving higher personal accuracy than a one-size-fits-all FL model, thus better handling individual expression variations.

2.3. Cross-Silo Federated Learning

This method works with a handful of stable, trusted clients like universities or big companies. In this setup, several well-equipped research labs, each with a high-quality but limited dataset of facial expressions, collaborate.

Li et al. combined FL with Meta-Learning (learning to learn) [13]. The whole FL process isn't just about training a model. It trains a model that can be quickly adapted (fine-tuned) with very little new data. The achievement is a highly efficient framework ideal for scenarios where a new research lab joins the consortium. The new lab can quickly customize the global model it learned earlier to fit its own data, without needing to restart everything from zero. This saves time, reduces the amount of new data and computing power needed, and keeps everyone's privacy intact across all the different sites.

2.4. Cross-Device Federated Learning with Limited Data

This is the most challenging setting, involving millions of unreliable devices (e.g., smartphones) with limited data per device. A hypothetical but research-grounded application is described in Huang et al.'s papers [14]. The data could be short video clips of users consenting to train an emotion-aware keyboard. Each user's device trains a lightweight model on a handful of personal examples. The central server aggregates updates from thousands of devices. The key achievement is the engineering feat of handling extreme scalability, device dropout, and non-IID data (each user's emotional data is unique). The new global model makes it easier for everyone to recognize emotions better, all while keeping personal video clips private on their own devices. It's designed to protect your data and make sure your videos never leave your device.

2.5. Hybrid CNNs for Multimodal Federated Learning

Human emotions are complex, and relying solely on facial images can sometimes make accurate judgment difficult. Multimodal learning improves the accuracy and robustness of emotion recognition by combining multiple sources of information (such as visual/facial, audio/tonal, and text/linguistic).

Feng et al. proposed a framework [15] where different modalities (e.g. visual/facial and audio) remain on different clients. Their model uses a CNN-based network (like ResNet) to extract features from facial images on one client and a separate network for audio on another. The key achievement is a novel fusion module that allows the global model to learn joint representations without ever sharing the raw facial or audio data. This method makes the emotion classification system much more reliable by using different clues from multiple senses, all while keeping privacy secure through a strict federated learning setup.

3. Discussion

3.1. Challenges

3.1.1 Data Heterogeneity (Non-IID Data)

In facial emotion classification, data across clients (e.g., different hospitals, user groups from various regions) are inherently non-IID. This manifests as feature distribution skew (clients have images of people with different ethnicities, ages, and lighting conditions) and label distribution skew (certain emotions may be more frequently expressed or labeled in one culture than another). The usual Federated Averaging (FedAvg) algorithm doesn't work as well when the data across devices isn't similar or evenly distributed.

Hsu et al. demonstrates that as the degree of non-IIDness increases, the accuracy of models trained with FedAvg drops precipitously [16]. In emotion recognition, this is exacerbated by the subjective nature of emotion labels themselves, which can vary between annotators from different backgrounds, further complicating the learning process and leading to a model that is biased towards the majority or most represented client groups.

3.1.2 The Privacy-Performance Trade-off

Even though federated learning (FL) stops raw data from being directly shared, recent research shows that the updates shared during training—like gradients—can sometimes be turned back into the original sensitive data. This kind of security flaw has been seen in attacks such as Deep Leakage from Gradients. For facial emotion data, this is a critical flaw. An adversarial server could potentially reconstruct a user's face from the gradients sent during training, completely undermining FL's privacy promise.

To mitigate this, techniques like Differential Privacy (DP) are employed. However, adding noise to gradients for differential privacy ends up hurting the model's performance [17]. In emotion classification, small facial movements are really important. When trying to protect privacy by adding noise, it often makes it harder for the model to pick up those tiny cues. This can cause the model's accuracy to drop a lot. This creates a direct tension between the level of privacy protection and the final mode's performance, a challenge that current methods have not yet satisfactorily resolved.

3.1.3 System and Statistical Heterogeneity

Real-world FL systems must contend with variability in the hardware and connectivity of client devices (system heterogeneity). Some devices may be powerful servers, while others are resource-constrained mobile phones. Furthermore, the amount of data per client can vary dramatically (statistical heterogeneity). A standard FedAvg protocol can be stalled by "stragglers"—slow clients that delay the global aggregation process.

Wang et al. attempts to address architectural alignment, but the core issue remains [18]. Clients may drop out mid-training, and forcing a one-model-fits-all approach can lead to inefficiency and exclusion of valuable but resource-poor data contributors, biasing the model towards data from users with more advanced hardware.

3.2. Future prospects

3.2.1 Personalized and Semi-Supervised Federated Learning

Instead of seeking a single global model, a major future direction is to create personalized models that adapt to each client's unique data distribution. Techniques such as multi-task learning [19] treat each client's learning framework as related but independent tasks, thereby allowing the model to consider individual or cultural differences in emotional expression.

Given the high cost of labeling facial emotions, combining FL with semi-supervised learning is a promising avenue. Clients could use a small amount of labeled data and a large corpus of unlabeled personal images to train locally. A semi-supervised federated learning [20] setup might work well for emotion classification. It can cut down the amount of labeling needed from users, while still making use of the lots of unlabeled data they have. This way, the model can learn better and be more flexible across different situations.

3.2.2 Cross-Modal and Transfer Learning

Emotion is not expressed solely through facial visuals. It is multimodal, encompassing speech, text, and physiological signals. Future FL systems for affective computing will likely move towards cross-modal FL. Han et al. discuss how it's possible to train models [21] where one client provides visual data, another supplies audio, and a third shares text, all without having to share the actual raw data. A global model could learn a joint representation space, enriching the emotion classification capability beyond what any single modality could achieve.

At the same time, the FTL was also employed. A common practice is to fine-tune a pre-trained model (e.g. a face recognition network like FaceNet) within the FL framework. This method called "Federated Fine-Tuning," which is used with models pre-trained on big datasets like AffectNet or FER, can help the model learn faster and work better. This is especially useful when there's not a lot of data from the clients. The pre-trained model provides a strong feature extraction foundation, and the FL process specializes it for emotion in a privacy-preserving manner.

3.2.3 Blockchain-Enhanced and Trustworthy Federated Learning

To address trust issues, future systems may integrate FL with blockchain technology to create a transparent and auditable training process. Smart contracts can help manage how clients are chosen and how incentives are given out, making sure everyone contributes fairly. More importantly, blockchain can combat poisoning attacks by providing an immutable record of model updates, making it easier to identify and reject malicious contributions.

Qin et al. suggests that a decentralized ledger can replace the central server, further enhancing the system's robustness and trustworthiness [22]. For sensitive tasks like recognizing emotions, it’s really important to have a clear, trustworthy, and decentralized way to track how the model was trained. This helps meet regulations and makes users more comfortable with the technology, especially when there's a risk of bias or someone trying to manipulate the system.

4. Conclusion

This review shows that FL could be a good way to protect privacy when classifying facial emotions. By enabling collaborative model training without raw data sharing, FL mitigates critical privacy concerns while maintaining competitive performance. Methods like horizontal, personalized, and cross-silo federated learning, along with architectures like CNNs and ResNet, have been effective in dealing with various types of emotional data that aren’t the same across sources. There are still a lot of challenges to get past, like dealing with different types of data, balancing privacy with performance, and working within system limits. Future work should prioritize personalized and semi-supervised FL, cross-modal integration, and blockchain-based trust mechanisms to enhance robustness, adaptability, and real-world applicability of FERC systems.

References

[1] Capriola-Hall N N. Group Differences in Facial Emotion Expression in Autism: Evidence for the Utility of Machine Classification. Behavior Therapy, 2019, 50 (4): 828-838.

[2] Metzen E, Nayyeri D M, Schäfer R, et al. Emotion Processing of Facial Affect Expression in Patients with Somatic Symptom Disorder with Predominant Pain - An EEG-Study. NeuroImage, 2025, 307: 121036.

[3] Michele P D, Vincenzo O, Giovanna F, et al. Differences in Facial Emotion Recognition Between Bipolar Disorder and Other Clinical Populations: A Systematic Review and Meta-Analysis. Progress in Neuro-Psychopharmacology & Biological Psychiatry, 2023, 127: 110847.

[4] Xiang G, Yao S, Wu X, et al. Driver Multi-Task Emotion Recognition Network Based on Multi-Modal Facial Video Analysis. Pattern Recognition, 2025, 161: 111241.

[5] Virolle J, Mouchet S, Robert L, et al. Facial Emotion Recognition in Sexual Offenders. Aggression and Violent Behavior, 2024, 78: 101982.

[6] Konečný J, McMahan H B, Yu F X, et al. Federated Learning: Strategies for Improving Communication Efficiency. arXiv, 2016.

[7] Abdulrahman S, Tout H, Ould-Slimane H, et al. A Survey on Federated Learning: The Journey from Centralized to Distributed On-Site Learning and Beyond. IEEE Internet of Things Journal, 2021, 8 (7): 5476–5497.

[8] Shome D, Kar T. FedAffect: Few-Shot Federated Learning for Facial Expression Recognition. 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2021: 4151–4158.

[9] Su X, Zhou Y, Cui L, et al. On Model Transmission Strategies in Federated Learning with Lossy Communications. IEEE Transactions on Parallel and Distributed Systems, 2023: 1–14.

[10] McMahan B, Moore E, Ramage D, et al. Communication-Efficient Learning of Deep Networks from Decentralized Data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), 2017: 1273-1282.

[11] Arivazhagan M G, Aggarwal V, Singh A K, et al. Federated Learning with Personalization Layers. arXiv, 2019.

[12] Fallah A, Mokhtari A, Ozdaglar A. Personalized Federated Learning with Theoretical Guarantees: A Model-Agnostic Meta-Learning Approach. Advances in Neural Information Processing Systems, 2020, 33: 3557-3568.

[13] Li X, Jiang M, Zhang X, et al. Federated Meta-Learning for Facial Emotion Recognition. Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020: 6522-6529.

[14] Huang Y, Du J, Yang Z, et al. Towards Federated Learning for Emotion Recognition from Mobile Device Sensors. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2021, 5 (3): Article 156.

[15] Feng Y, Li B, Wang Y, et al. FedMSA: A Benchmark for Federated Multi-Modal Sentiment Analysis with Cross-Modal Self-Attention. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024: 1-10.

[16] Hsu T M H, Qi H, Brown M. Measuring the Effects of Non-Identical Data Distribution for Federated Visual Classification. arXiv, 2019.

[17] Pan K, Ong Y S, Gong M, et al. Differential Privacy in Deep Learning: A Literature Survey. Neurocomputing, 2024, 589: 127663.

[18] Wang H, Yurochkin M, Sun Y, et al. Federated Learning with Matched Averaging. arXiv, 2020.

[19] Smith V, Chiang C K, Sanjabi M, et al. Federated Multi-Task Learning. Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017), 2017: 4424-4434.

[20] Jeong W, Yoon J, Yang E, et al. FedMatch: Federated Learning Over Heterogeneous Data. Proceedings of the 37th International Conference on Machine Learning, 2020: 5136-5145.

[21] Han S, Cheng H, Zhang Y, et al. FedMultimodal: A Benchmark for Multimodal Federated Learning. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022: 9232-9250.

[22] Qin Y, Wang S, Yi S, et al. Blockchain-Based Federated Learning: A Comprehensive Survey. International Journal of Intelligent Systems, 2021, 36 (11): 8000-8025.