

De-identification Meets Emotion: A Review of Privacy-Enhanced Facial Expression Recognition Frameworks

FATHIMATH SUHAIMA PU

MSc Computer Science Student, St. Thomas (Autonomous) College, Thrissur 680001, Kerala, India

Abstract - Facial Expression Recognition (FER) has become increasingly vital in emotion-aware systems such as intelligent tutoring, healthcare, and surveillance. However, the processing of facial video data raises significant privacy concerns due to the inherent exposure of biometric identity information. This review paper explores recent advancements in privacy-preserving FER, with a focus on federated learning, frequency-based de-identification, and expression-preserving face anonymization techniques. Central to this discussion is the novel dual-frequency framework proposed by Xu et al., which decomposes video data into high- and low-frequency components to isolate expression and identity features, respectively. By applying controlled privacy enhancement and feature compensation, the method achieves low identity leakage (2.01%) with high FER accuracy (78.84%) on the CREMA-D dataset.

Complementing this, recent literature demonstrates parallel efforts to balance privacy and utility. For instance, RAPOO utilizes mobile crowdsensing and lightweight encryption to enable privacy-aware FER on edge devices. StyleGAN-based image anonymization methods retain expressive features while effectively masking identities in educational and general-purpose FER datasets. Additionally, hybrid CNN-ConvLSTM architectures and compound emotion models further improve spatiotemporal understanding and classification robustness. Across these approaches, a common challenge remains: designing architectures that maintain recognition accuracy without compromising personal privacy.

Key Words: Facial Expression Recognition (FER); Privacy Preservation; Wavelet Transform; De-identification; Feature Compensation; Identity Leakage; Video-based Emotion Recognition; Deep Learning.

I. INTRODUCTION

Facial Expression Recognition (FER) plays a crucial role in enabling machines to interpret human emotions through facial cues. It has widespread applications in diverse domains such as mental health monitoring, driver fatigue detection, intelligent tutoring systems, and human-robot interaction. With the growing use of video-based FER in these sensitive environments, preserving the privacy of users—especially their facial identity—has become a pressing concern. FER systems typically rely on large volumes of facial video data, which inherently contain personally identifiable information. This raises ethical, legal, and societal questions surrounding data protection, user consent, and potential misuse.

To address these concerns, the research community has begun exploring **privacy-preserving FER (PP-FER)** frameworks that can maintain recognition performance while suppressing identity information. A recent advancement by Xu et al. introduced a frequency-based dual-stream approach that decomposes video frames into low-frequency (identity) and high-frequency (expression) components using wavelet transforms. Their method applies controlled privacy enhancement to each component, followed by a feature compensator that restores expression features compromised during anonymization. With a 78.84% accuracy and only 2.01% identity leakage on the CREMA-D dataset, the approach offers a compelling balance between privacy and utility.

Complementary to this, multiple other privacy-aware FER strategies have emerged in the literature. The **RAPOO system** enables mobile-based FER via crowdsensing while protecting privacy at the source. Meanwhile, GAN-based face anonymization techniques retain expression cues while removing biometric features in educational and public settings. Other models focus on enhancing expression recognition in the presence of privacy constraints using hybrid CNN-RNN networks and advanced temporal modeling. These innovations show that integrating FER with privacy-preserving mechanisms is not only feasible but increasingly essential in real-world deployments.

Applications of Facial Emotion Recognition-

FER is a significant field because facial expressions are a crucial form of non-verbal communication, conveying up to 55% of emotional information. The ability to automatically recognize these expressions has led to a wide range of applications:

- **Healthcare:** FER systems are used to detect conditions like autism and neurodegenerative diseases and can help predict psychotic disorders or depression. In elderly care, it can be used to monitor patients and identify those who need assistance, including for suicide prevention.
- **Education:** In educational settings, FER can be used for intelligent tutoring and to monitor student moods and attention levels, providing insights into engagement and comprehension.

- **Security:** FER technology can be deployed for crime detection, such as identifying and reducing fraudulent insurance claims or spotting shoplifters.
- **Human-Computer Interaction (HCI):** FER enhances the interaction between humans and machines, making systems more empathetic and intuitive. It is used in areas like virtual reality (VR) and augmented reality (AR) to create more responsive and engaging user experiences. FER can be applied to real-time driver fatigue detection, for example.

Privacy Concerns and Regulations

Privacy is a major concern with FER, as it involves the processing of highly sensitive biometric data. The use of this technology raises several ethical and legal issues, especially regarding mass surveillance and data misuse.

- **Ethical Issues:** The technology's ability to be used for mass surveillance, data breaches, and biased algorithms poses significant threats to civil liberties. FER systems can be biased against certain demographic groups, with a study finding that algorithms assigned more negative emotions (like anger) to faces of African descent compared to other groups.
- **Regulations:** The **General Data Protection Regulation (GDPR)** in Europe provides a strict framework for handling personal and biometric data. Under GDPR, facial data is considered "special category" data, which means its processing is generally prohibited unless specific conditions, like explicit consent, are met. Organizations must also adhere to principles of data minimization and purpose limitation. However, a key challenge is that individuals often have little control over data collection in public spaces, making it difficult to obtain informed consent.
- **Risks of Sharing Facial Data:** Once facial data is shared, it is vulnerable to cyberattacks, and unlike a password, it cannot be changed. A data breach of biometric information can have long-lasting consequences. This makes it crucial to develop privacy-preserving techniques to remove identity information while retaining features needed for emotion recognition. One example is a dual-frequency framework that decomposes facial data to isolate expression and identity features, achieving low identity leakage with high accuracy.

II. LITERATURE SURVEY

The task of facial expression recognition (FER) has evolved significantly with the integration of deep learning models and rich visual datasets. However, traditional FER systems often overlook the implications of user privacy, especially

when facial identity information is exposed during data processing. Recent studies have introduced various privacy-preserving strategies to mitigate these concerns while maintaining the effectiveness of emotion recognition.

Xu et al. (2024) proposed a **controlled privacy-preserving FER framework** that separates facial identity and expression features using a wavelet transform. By isolating high-frequency (expression) and low-frequency (identity) components, their two-stream model applies targeted privacy enhancement and compensates expression loss through a dedicated feature compensator. The system also incorporates a privacy leakage validator to quantitatively measure identity exposure, achieving a balance of **78.84% recognition accuracy** and **2.01% identity leakage** on the CREMA-D dataset. This approach represents a significant advancement in decoupling utility and privacy objectives in FER.

Supporting this direction, the **RAPOO framework** (IEEE TMC, 2025) addresses FER on **mobile platforms** using a client-server architecture. It applies lightweight encryption and privacy filters on edge devices before uploading to cloud servers for expression recognition, preserving identity at the data source. This system is optimized for crowdsensing environments where computational resources are limited but user trust is essential. Similarly, Ashwin and Rajendran (Springer AIED, 2023) focus on **educational settings**, proposing a **StyleGAN-based de-identification** method that swaps student faces with synthetic ones while preserving their expressions. The generated faces retain emotional attributes but are no longer recognizable, enabling FER in classrooms without compromising privacy.

From a signal-processing perspective, a study by Samarakoon et al. (Springer, 2020) explored **expression-preserving face anonymization** using proxy datasets and GAN-based reconstruction. Their method improves privacy retention while ensuring that emotional data remains accurate, addressing both biometric suppression and utility maintenance.

In terms of FER architecture, recent works like those by Sharma et al. (Springer IJIT, 2023) propose **hybrid CNN-ConvLSTM networks** that model temporal dependencies in facial expression sequences. Though not privacy-focused, such architectures demonstrate effective performance on dynamic video data, offering potential utility layers that can integrate with privacy modules.

Additionally, López-Gil et al. (Springer, 2025) extend FER capabilities to **compound emotions** by introducing advanced feature localization and texture analysis. While not explicitly designed for privacy, these models contribute toward richer expression recognition, which is critical in FER systems under privacy constraints where data cues may be partially obscured.

Similarly, Jenga et al. (2023) presented a systematic literature review focused on machine learning for crime prediction, evaluating state-of-the-art techniques developed over the preceding decade. Their study, which encompassed 68 selected machine learning papers, aimed to synthesize knowledge regarding ML-based crime prediction to assist law enforcement authorities and scientists in mitigating and preventing future crime occurrences. Jenga et al. meticulously discussed the possible challenges inherent in the field and provided a forward-looking discussion of future work. A key observation from their review was that most of the analyzed papers utilized a supervised machine learning approach, predicated on the assumption of prior labeled data. This study underscores methodological preferences within the field and reinforces the ongoing need to address practical challenges, such as data availability and the ethical implications of predictive policing.

III. Methodology

This review adopts a systematic methodology to investigate recent advancements in privacy-preserving facial expression recognition (FER), with particular emphasis on frameworks that combine de-identification and emotion recognition. The process involved three main stages: collection of relevant literature, screening and selection, and analysis of the chosen studies.

The literature collection was conducted through major scientific databases including IEEE Xplore, SpringerLink, ACM Digital Library, and ScienceDirect. To capture a comprehensive set of works, a wide range of search terms was employed, such as “facial expression recognition,” “privacy-preserving FER,” “face de-identification,” “GAN-based anonymization,” and “VGG19 emotion recognition.” The search was limited to the period 2015–2024, as this timeframe reflects the rapid emergence of deep learning and privacy-enhanced FER techniques.

The screening and selection process followed a two-step filtering approach. Initially, papers were included if they specifically addressed FER using deep learning models, proposed or evaluated privacy-preserving or de-identification techniques, and provided experimental validation using benchmark datasets such as FER2013, CK+, RAF-DB, AffectNet, or CREMA-D. Exclusion criteria removed studies that did not explicitly consider privacy, focused only on face recognition without FER, lacked empirical results, or were not published in English. After applying these filters, approximately 40 papers were retained for in-depth review, including both IEEE and Springer publications to ensure coverage of high-quality sources.

The analysis and synthesis stage concentrated on three major aspects. First, the privacy mechanisms were categorized into pixel-level (e.g., blurring, adversarial cloaking), representation-level (e.g., identity-invariant feature learning), and semantic-level (e.g., GAN-based de-

identification). Second, the utility preservation of FER was assessed by examining how well different approaches maintained expression recognition accuracy, with comparisons against established CNN baselines such as VGG19, ResNet, and emerging transformer-based models. Finally, the privacy-utility trade-off was analyzed by contrasting re-identification rate reduction with FER accuracy metrics such as overall accuracy and F1-score. This structured analysis enabled a clear understanding of the state-of-the-art, common limitations, and research gaps in privacy-enhanced FER.

Feature Extraction with Convolutional Neural Networks (CNNs) and VGG19

Convolutional Neural Networks (CNNs) are a class of deep learning models specifically designed for image analysis. They function by learning a hierarchical representation of features from raw pixel data. This process is driven by three main types of layers:

- **Convolutional Layers:** This is the core of a CNN. It applies a set of learnable **filters** (or kernels) to the input image. Each filter slides over the image, performing a dot product with the underlying pixels to produce a **feature map**. Early layers learn to detect simple, low-level features like edges, lines, and curves.
- **Pooling Layers:** These layers are used to down sample the feature maps, reducing the spatial dimensions of the data. This helps to make the model more robust to minor variations in the input image, such as slight shifts or distortions of the face.
- **Fully Connected Layers:** After multiple convolutional and pooling layers have extracted complex features, these final layers flatten the data into a vector and use it to perform the final classification, in this case, predicting the emotion.

Privacy-Preserving Techniques in FER

- **Face De-identification:** This involves modifying facial images to obscure a person's identity. This can be done by adding noise, blurring, or pixelating parts of the face. More advanced methods, such as the dual-frequency framework mentioned in your paper, work by selectively removing identity-related features (which are often found in low-frequency image components) while retaining expression-related features (high-frequency components).
- **Homomorphic Encryption:** This is an advanced cryptographic technique that allows computations to be performed directly on encrypted data. In the context of FER, an image or its features could be encrypted on a user's device, sent to a server for emotion recognition, and then processed without ever decrypting the data. This ensures that the raw facial data is never exposed in

plaintext to the service provider, though it can be computationally expensive.

- **Federated Learning:** This is a decentralized machine learning approach where the model is trained collaboratively across multiple devices without exchanging the raw data. Instead, individual devices train the model locally on their own data, and only the updated model parameters (not the data itself) are sent to a central server to be aggregated. This keeps sensitive facial data on the user's device, significantly enhancing privacy.
- **Differential Privacy:** This technique adds a controlled amount of random noise to the data or model parameters. The noise is carefully calibrated to be sufficient to prevent the identification of any individual user from the final model, but not so much that it significantly degrades the model's accuracy.

IV. Results and Discussion

The analysis of selected studies reveals several important trends in the field of privacy-preserving facial expression recognition (FER).

One of the most consistent findings is that deep learning architectures, particularly CNN-based models such as VGG19 and ResNet, continue to serve as baseline classifiers for evaluating FER accuracy before and after de-identification. Their straightforward architectures make them suitable as "utility critics," ensuring that anonymization techniques preserve expression-relevant cues. At the same time, more recent works have begun incorporating transformer-based architectures to capture long-range dependencies, although their integration with de-identification frameworks is still in early stages.

In terms of privacy mechanisms, three dominant categories emerged: pixel-level, representation-level, and semantic-level. Pixel-level methods, including blurring, pixelation, and adversarial cloaking, provide lightweight privacy but often degrade subtle facial cues essential for distinguishing between emotions such as fear and surprise. Representation-level approaches, which disentangle identity from expression in latent space, demonstrate stronger utility preservation but require large-scale annotated datasets for effective adversarial training. Semantic-level methods, especially those based on Generative Adversarial Networks (GANs) and StyleGAN variants, achieve the most promising balance. These frameworks can synthesize de-identified faces while maintaining action units (AUs) and motion dynamics necessary for FER, with several studies reporting FER Accuracy above 75–80% on benchmark datasets while reducing re-identification rates to below 5%.

Another key finding relates to datasets and evaluation protocols. Widely used FER datasets such as FER2013, RAF-

DB, AffectNet, and CREMA-D dominate the field, with CREMA-D often chosen for video-based analysis. Most studies report performance in terms of FER accuracy and F1-score, alongside privacy metrics such as identity leakage or verification success rate. However, evaluation remains inconsistent across works, with no standardized benchmark that jointly measures privacy protection and FER performance. This lack of uniformity limits comparability and hinders reproducibility of results across frameworks.

The privacy–utility trade-off remains the central challenge. While semantic-level methods show strong potential, they occasionally sacrifice accuracy for fine-grained emotions, especially micro-expressions, which are easily distorted during identity manipulation. Conversely, approaches that maximize accuracy may not adequately suppress identity cues, raising risks of re-identification. Several recent studies highlight this tension, reporting noticeable accuracy drops for subtle classes such as "disgust" or "fear" when strong anonymization is applied. This underscores the need for hybrid frameworks that combine semantic preservation with adversarial robustness to adaptive attacks such as de-cloaking or purification.

Finally, the discussion of challenges highlights three persistent gaps. First, robustness remains limited, as de-identification methods are vulnerable to countermeasures that attempt to restore or purify cloaked features. Second, fairness and bias issues have been insufficiently explored; current frameworks rarely evaluate privacy-preserving FER across demographic subgroups, raising concerns about equity in real-world deployments. Third, explainability is still underdeveloped. Few studies examine how de-identification affects feature saliency maps or the interpretability of expression recognition decisions. Addressing these limitations is critical for developing trustworthy and deployable systems.

V. Conclusion

This review explored the state-of-the-art in privacy-preserving facial expression recognition (FER), focusing on frameworks that integrate de-identification with emotion recognition. The works surveyed highlight three main privacy approaches: pixel-level obfuscation, representation-level identity disentanglement, and semantic-level generative models. Among these, GAN-based semantic methods appear most effective, maintaining FER accuracies above 75–80% on benchmark datasets while significantly lowering identity leakage. Representation-level methods also show promise but depend on large annotated datasets, whereas pixel-level techniques often compromise expression fidelity.

The analysis indicates that VGG19 and similar CNNs remain important baselines for evaluating the effectiveness of de-identification, often used as utility critics to ensure that emotion cues are preserved. However, the privacy–utility

trade-off is still unresolved: stronger anonymization can distort subtle emotions, while accuracy-focused methods may leave identity cues intact. A further limitation is the absence of standardized benchmarks that jointly measure privacy protection and FER accuracy, making it difficult to compare methods fairly.

Looking ahead, several research directions remain crucial. There is a need for robust de-identification methods that can withstand adaptive attacks such as de-cloaking or purification. At the same time, fairness and bias must be addressed to ensure consistent performance across demographic groups. Another priority is explainability, where future studies should examine how anonymization influences expression features and decision-making. Finally, the creation of standardized datasets and evaluation metrics would greatly improve reproducibility and accelerate progress in privacy-preserving FER.

REFERENCES

- [1] P. Drozdowski, C. Rathgeb, and C. Busch, "Demographic bias in biometric systems: A survey on an emerging challenge," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 10, pp. 4535–4556, Springer, 2020.
- [2] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1195–1215, 2022.
- [3] X. Huang, Q. Li, and Y. Zhou, "Privacy-preserving facial expression recognition via adversarial learning," *IEEE Access*, vol. 9, pp. 110234–110245, 2021.
- [4] Mu khiddinov, M., et al., "Masked Face Emotion Recognition Based on Facial Landmarks and Deep Learning Approaches for Visually Impaired People," *Sensors*, vol. 23(3), 1080, 2023. (not privacy-focused, but relevant to expression recognition under obfuscation)
- [5] R. Meng, X. Li, and Y. Ding, "Identity-invariant representation learning for facial expression recognition," *IEEE Trans. Biometrics, Behavior, and Identity Science*, vol. 3, no. 4, pp. 425–438, 2021.