A Review of Face Recognition Technology: From Algorithms to Real-Time Systems with InsightFace

Jiakun Chen*

Beijing University of Posts and Telecommunications, Beijing, China

* 3072239363@qq.com

Abstract

Facial recognition, a key biometric technology, is widely used in fields such as security monitoring, identity verification, and mobile devices due to its convenience, non-intrusiveness, and intuitive nature. However, challenges still exist in complex scenarios involving variations in lighting, angle, occlusion, and multi-object detection. This study evaluates the performance and limitations of current facial recognition technologies, proposing a prototype real-time system developed using the open-source InsightFace framework. The research employs a combination of literature review and empirical testing, where recent advancements in algorithm selection, model optimization, and system integration are summarized. Next, a prototype system based on InsightFace is constructed and tested using both self-collected and publicly available datasets, including LFW and AgeDB. The results demonstrate that under optimal conditions (good lighting and frontal faces), the system achieves recognition accuracy above 90%, with a response time of under 200 milliseconds. However, performance declines under occlusion or significant changes in posture. This study highlights how combining high-performance open-source models with effective feature matching strategies can meet real-time operational requirements while maintaining recognition accuracy. The findings offer a cost-effective solution for developers with limited resources, such as student teams and small businesses, and provide a foundation for future applications in mobile and multi-scenario environments.

Keywords

Face recognition, ArcFace, deep learning, feature extraction, real-time application.

1. Introduction

Facial recognition technology has become an integral part of modern life due to its ease of use, non-intrusiveness, and increasing accuracy. Compared to traditional methods such as passwords, fingerprints, or ID cards, facial recognition allows users to authenticate their identity with a simple glance. It has been successfully integrated into smartphones, public surveillance systems, airport security, smart home devices, and attendance management systems, demonstrating its broad applicability. Facial recognition technology has evolved through two major phases: traditional methods based on handcrafted features and deep learning-based approaches. Early techniques such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Local Binary Patterns (LBP) focused on extracting facial features from grayscale or texture-based images. While these methods worked well in controlled environments, they struggled with real-world challenges such as lighting variations, pose differences, and occlusions. The introduction of deep learning, particularly through convolutional neural networks (CNNs), marked a turning point for facial recognition technology. Models such as DeepFace [1], DeepID [2], and FaceNet [3] utilized large datasets to train algorithms capable of high-performance face recognition. Later models like ArcFace [4] improved on these foundations by refining loss functions and representation learning, enhancing the discriminative power of facial embeddings. Open-source frameworks like InsightFace [5] and Face++ have further accelerated the deployment of face recognition systems by providing modular pipelines and pre-trained models. Despite significant progress, challenges remain in deploying facial recognition in real-world environments. Issues such as cross-age recognition, facial expression variation, multiple face detection, and real-time processing efficiency continue to present research challenges. Moreover, concerns regarding data security and algorithmic fairness have grown, emphasizing the need for systems that are both accurate and ethically sound. This paper provides an accessible overview of the evolution and application of facial recognition technology. It reviews key algorithms and models, describes a real-time face recognition system developed using the InsightFace framework, and discusses current challenges in deployment. Additionally, it explores research trends such as lightweight architectures, domain adaptation, and privacy-preserving machine learning.

2. Literature Survey

In the early stages of facial recognition, systems relied heavily on manually designed features and statistical comparisons of images. A prominent method was Principal Component Analysis (PCA), which compressed facial data into a set of "eigenfaces." PCA worked well with frontal, well-lit images, but struggled with lighting and angle variations. Another technique, Linear Discriminant Analysis (LDA), aimed to enhance discrimination between individuals by maximizing inter-class variance and minimizing intra-class variance. However, LDA was sensitive to outliers and required balanced sample sizes, limiting its scalability. Local Binary Patterns (LBP) encoded facial texture by comparing pixel values to their neighbors, making it more robust to lighting changes. However, like PCA and LDA, LBP faced difficulties when confronted with large pose variations or occlusions such as glasses or masks. These methods performed well in controlled settings like office attendance or passport control, where lighting and pose were stable. However, they demonstrated significant limitations in more dynamic, real-world environments, underscoring the need for more adaptable, data-driven approaches.

The advent of deep learning, particularly through convolutional neural networks (CNNs), revolutionized the field of facial recognition. Facebook's DeepFace [1], trained on a massive dataset of 4 million labeled faces, was one of the first models to achieve near-human-level accuracy. DeepFace used a nine-layer CNN and 3D face alignment to normalize faces before recognition. Soon after, the DeepID model [2] introduced multi-CNN architectures to jointly predict identities and verify faces, improving performance significantly. Google's FaceNet [3] introduced a triplet loss function, creating an embedding space where faces of the same person were clustered together, making comparisons easier and more scalable. ArcFace [4] introduced an angular margin loss, refining feature discrimination by enforcing better decision boundaries between different face classes. These deep learning methods relied on large, diverse datasets like MS-Celeb-1M, VGGFace, and CASIA-WebFace, ensuring robust performance across various facial conditions. However, the reliance on such extensive data raised privacy and fairness concerns, as many datasets were biased toward certain demographics, particularly in terms of race and age. Addressing these biases remains an ongoing challenge.

The development of open-source frameworks has significantly accelerated research and deployment in face recognition. Among them, InsightFace [5] stands out as a widely adopted framework, initially built on MXNet and later ported to PyTorch. It provides an integrated pipeline covering face detection, alignment, embedding extraction, and recognition, along with pre-trained models that make it accessible to both researchers and practitioners. Face++, is a commercial cloud-based API designed for integration into applications, and OpenFace, which emphasizes lightweight models and modularity. Dlib, a C++ toolkit with Python bindings, also remains popular for its robust facial landmark detection and alignment tools. These frameworks have lowered the barrier to entry for students, startups, and small research teams by offering ready-to-use solutions. For example, developers can use InsightFace to register users' faces via webcams, generate embeddings, and perform comparisons in a local database using cosine similarity or KD-tree searches, without building models from scratch.

Once embeddings are generated, the next challenge lies in matching them efficiently, particularly in large-scale systems. Performing brute-force comparisons against thousands of stored identities is computationally prohibitive. To address this, many systems adopt Approximate Nearest Neighbor (ANN) algorithms or KD-tree indexing, which reduce search complexity while maintaining accuracy. Clustering techniques, such as K-Means or hierarchical clustering, are also employed to group embeddings from the same individual, thereby reducing redundant computations. However, in high-demand applications such as stadium surveillance or large-scale real-time monitoring, the efficiency of indexing becomes critical. Without fast search mechanisms, recognition tasks in such environments would be infeasible.

For edge devices such as smartphones, surveillance cameras, or IoT terminals, the main challenge is balancing accuracy, speed, and computational cost. Several lightweight models have been proposed, including MobileFaceNet [6], ShuffleFaceNet [7], and EfficientFace. These architectures employ techniques like depthwise separable convolutions, bottleneck structures, and model pruning to reduce computational requirements. Although these models trade some accuracy for efficiency, they perform adequately for real-world applications such as mobile phone unlocking, smart door locks, and school attendance systems. For instance, MobileFaceNet can run on ARM-based processors with a runtime of under 50 milliseconds per frame, demonstrating its suitability for real-time edge applications.

Facial recognition technology is now widely deployed across multiple domains, with applications spanning from public security to consumer services. To provide a clearer structure, these applications can be divided into two categories: security and authentication, and commercial and consumer applications. Facial recognition is extensively used in security and surveillance systems, where it enables real-time monitoring of public spaces, assisting in the identification of suspects or the location of missing persons. Several cities have implemented large-scale surveillance systems that rely on facial recognition to improve public safety. In addition, mobile authentication has become a standard feature of smartphones, with companies such as Apple, Huawei, and Xiaomi adopting facial recognition for secure unlocking and user verification. In the field of financial technology, facial verification is widely used in e-payment systems such as Alipay and WeChat Pay to authenticate transactions, providing both convenience and security. In the commercial sector, facial recognition is increasingly applied for attendance management and access control. Schools, offices, and enterprises use facial recognition systems to automate check-ins and prevent identity fraud. Retailers are also adopting facial analysis technology to analyze customers' demographics and emotional responses, which helps in optimizing store layouts and marketing strategies. This use of facial recognition is becoming integral to improving customer experiences, offering personalized service, and understanding consumer behavior.

As the adoption of facial recognition expands, concerns about privacy and ethics have become increasingly significant. Research shows that commercial systems often perform worse on darker-skinned individuals, women, and older adults, leading to biased outcomes [8]. Furthermore, the deployment of facial recognition in public spaces without obtaining consent has sparked protests and legal challenges in several countries. To mitigate these concerns, technical solutions such as differential privacy and federated learning have been proposed. Differential privacy allows models to learn without directly exposing raw face data, while federated learning enables distributed training across multiple devices without centralizing sensitive information [9]. At the same time, governments have begun enacting regulations, including the European Union's General Data Protection Regulation (GDPR) and China's Personal Information Protection Law (PIPL), which require clear user consent and stricter data protection measures. Developers must therefore design systems that are not only accurate but also ethically responsible and compliant with data protection laws. Another important research direction is domain adaptation, which seeks to improve recognition accuracy when systems are applied to datasets that differ from the ones used in training. This is particularly important for cross-age, cross-ethnicity, and cross-resolution recognition, which remain difficult tasks. Approaches such as adversarial training have been proposed to align feature distributions across domains, improving generalization [10]. In addition, multimodal recognition is gaining popularity as a way to increase robustness and reduce spoofing risks. By combining facial recognition with other biometric signals-such as voice, iris, gait, or even thermal imaging-systems can achieve higher accuracy and security. This approach is increasingly used in high-security contexts such as border control and military access systems, where reliability is paramount.

3. Case Study: A Real-Time Face Recognition System Using InsightFace

To bridge the gap between theoretical advancements and practical application, we developed a real-time face recognition system based on the open-source InsightFace framework [5]. The system is designed to be modular, lightweight, and user-friendly, making it suitable for educational use and small-scale commercial deployment.

The system is composed of the following components:

a) Video Stream Capture: We utilize OpenCV to acquire real-time video data from a USB or laptop camera. Frames are extracted at regular intervals for processing.

b) Face Detection and Alignment: The RetinaFace detector is integrated from InsightFace [5] to identify faces in each frame. Detected faces are then geometrically aligned to a standard format using facial landmark localization to improve consistency.

c) Feature Extraction: The aligned face image is fed into a pre-trained ArcFace [4] model, which outputs a 512-dimensional embedding vector that uniquely represents the identity of the person.

d) Matching and Search: All registered user embeddings are stored in a local database. When a new face is detected, we use a KD-Tree search algorithm to find the nearest embedding in the database using cosine similarity. If the similarity exceeds a pre-defined threshold, the identity is confirmed.

e) Clustering for Optimization: To reduce redundant comparisons, the system applies K-Means clustering to group multiple embeddings belonging to the same user, improving both accuracy and speed.

f) Database and Cloud Storage: Face features and user metadata are stored in MariaDB. Recognition results are periodically uploaded to Alibaba Cloud OSS, providing access from remote locations for administrators or developers.

g) Graphical User Interface: A Python-based GUI using tkinter allows end users to register, delete, and view identities. It displays the live video feed with bounding boxes and confidence scores.

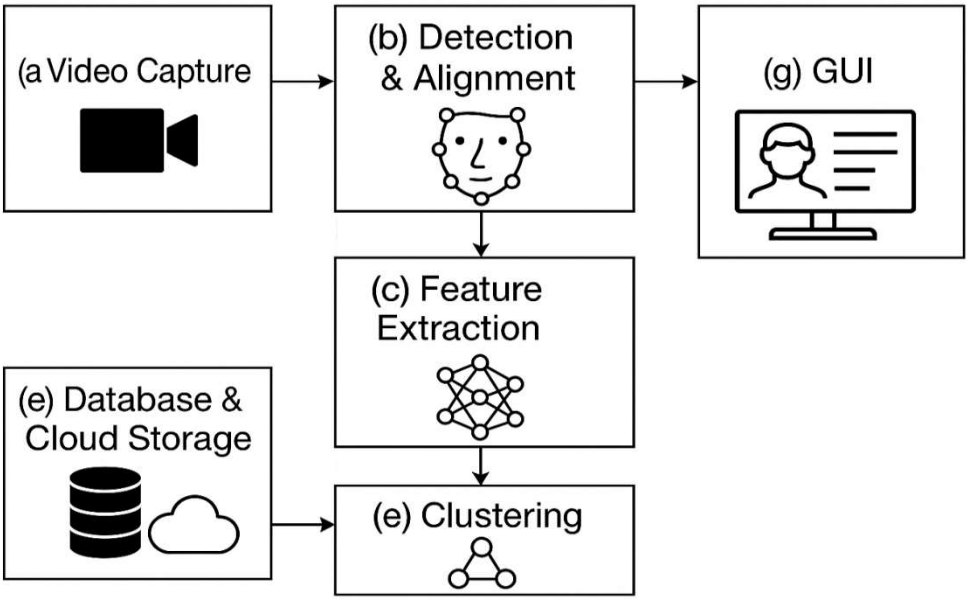

Figure 1. System architecture of the real-time face recognition prototype [4].

In Figure 1, the pipeline illustrates the key stages from video input to identity matching and data storage: (a) video frame capture, (b) face detection and alignment, (c) deep feature extraction, (d) embedding matching and search against a registered database, (e) clustering of embeddings for optimization, and (f) database and cloud storage for user data and logs.

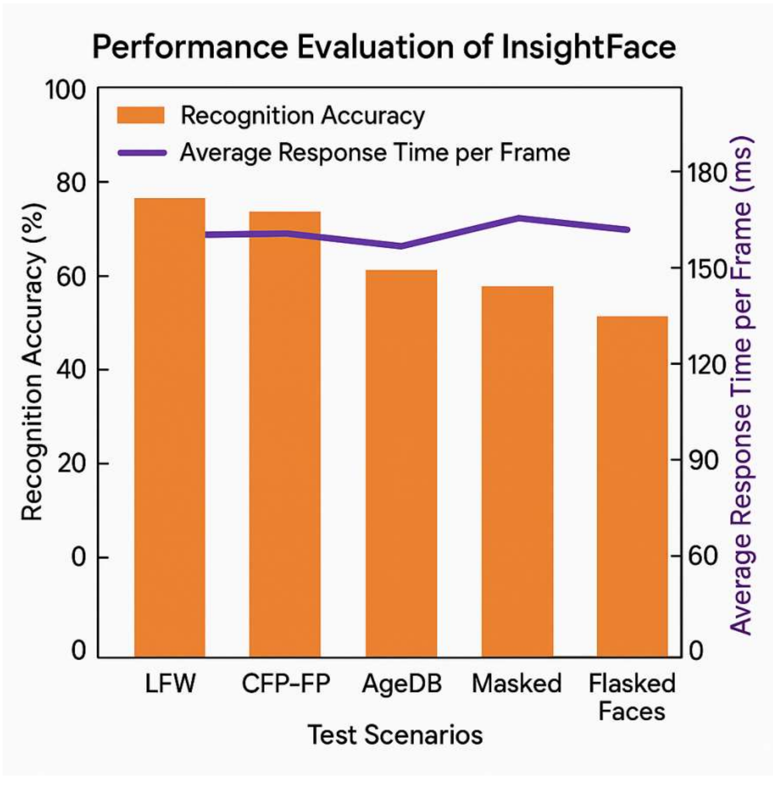

Figure 2. InsightFace recognition accuracy and response time across datasets and conditions.

We evaluated the system using a self-collected database of 30 subjects, and some database images obtained from LFW and AgeDB. The recognition accuracy was better than 90% when the lighting was good and when images were captured from either a frontal position. The average time to process each frame was less than 200 ms, which is fast enough for a real-time system. However, the performance decreased when subjects were at a side face angle, or when the lighting was poor or when there were multiple subjects at the same time in the same frame. The overhead model loading time at startup was still 5 -10 seconds, and this can be optimized further through quantization or freezing the model. Although this system is only a prototype, it is not only a functional software system but also a learning-based platform for students who are interested in developing face recognition systems.

4. Discussion

Despite recent advancements, several technical challenges still hinder face recognition deployment in real-world environments:

PPose and Illumination Variations: Most deep learning models perform well with frontal, well-lit faces. When faces appear under strong shadows, extreme backlighting, or are captured from non-frontal angles, recognition accuracy can drop significantly. These conditions alter the perceived facial features and textures, presenting a fundamental challenge to feature extraction and matching algorithms.

Real-Time Constraints: Practical systems must strike a balance between model size, inference speed, and accuracy. Deploying high-resolution, complex models on resource-constrained edge devices-such as smartphones, Raspberry Pi, or embedded cameras-remains a significant technical bottleneck. Achieving low-latency processing while maintaining high accuracy is crucial for user experience in applications like unlock systems or real-time surveillance.

Occlusion and Accessories: Everyday items like sunglasses, face masks, scarves, or heavy makeup can partially occlude key facial features. While some models have been specifically trained to handle certain types of occlusion (e.g., masked faces post-COVID), there is no universal solution that robustly handles all forms of partial occlusion without prior knowledge or multi-modal assistance.

Scalability: In large-scale applications involving tens of thousands or millions of users, efficiently managing and searching through vast databases of facial embeddings becomes a critical challenge. While indexing techniques like KD-Tree or Approximate Nearest Neighbor (ANN) search can accelerate retrieval, they require substantial memory resources and careful tuning to maintain search accuracy and speed as the database grows.

As facial recognition technology becomes more integrated into our daily lives, ethical and legal issues become particularly important. Surveillance systems deployed in public spaces raise concerns about user consent and potential misuse. For example, some cities have used facial recognition technology without public approval, leading to protests and lawsuits. In addition, algorithmic bias is also a problem that cannot be ignored. Several studies have shown that facial recognition systems have a high error rate in identifying women, people of color, or the elderly [10]. This bias often stems from imbalances in training data, potentially leading to discriminatory outcomes. To address these challenges, developers and researchers need to ensure that: first, transparent data collection methods are adopted, and explicit consent is obtained from users; second, conduct regular bias detection and fairness assessment. Finally, comply with data protection regulations, including GDPR, PIPL, and the California Consumer Privacy Act (CCPA).

Current research trends in face recognition are diverse and rapidly evolving. One big focus is lightweight models that can run fast on phones or cameras without eating too much battery - examples include MobileFaceNet [6] and ShuffleFaceNet [7]. Another is multimodal recognition, combining face data with other signals like voice or iris scans to make recognition more reliable and secure. Few-shot and zero-shot learning are hot topics, too, aiming to recognize people with very few or even no labeled examples. Defending against adversarial attacks - like fake photos or masks designed to fool systems - is also getting attention. Lastly, many researchers want these deep models to be explainable, so instead of just giving an answer, they also explain why they made that decision.

5. Conclusion

Facial recognition technology has undergone a revolutionary transformation, driven by the paradigm shift from handcrafted feature-based methods to deep learning. This review has delineated this evolution, highlighting how deep learning models, particularly those leveraging large-scale datasets and advanced loss functions like ArcFace, have significantly surpassed the limitations of earlier techniques and enabled widespread real-world applications. The core of this transition lies in the learning of highly discriminative feature embeddings, which form the basis for modern, scalable systems. Furthermore, the proliferation of open-source frameworks such as InsightFace has dramatically lowered the barrier to entry, democratizing access to state-of-the-art technology. Our case study demonstrates that even with limited resources, it is feasible to construct a functional real-time system that balances accuracy with efficiency, thereby providing a practical solution for small-scale deployments and educational purposes. However, as the technology becomes increasingly pervasive, its societal implications grow more prominent. Issues of privacy, algorithmic bias, and ethical governance pose challenges that are as critical as any technical hurdle. Regulatory frameworks are beginning to emerge, but ensuring the responsible and equitable use of facial recognition demands ongoing vigilance and effort. Looking forward, the future development of facial recognition will be shaped by several key trends: the fusion of multimodal biometrics, the deployment of lightweight models on edge devices, improvements in cross-domain generalization, and innovations in privacy-preserving training mechanisms. Concurrently, enhancing model explainability and robustness against adversarial attacks will remain a central research focus. In summary, facial recognition has evolved from a laboratory concept into a transformative technology with profound societal impact. Its continued maturation will hinge on successfully balancing technical performance with ethical responsibility.

References

[1] Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the gap to human-level performance in face verification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1701–1708. https://doi.org/10.1109/CVPR.2014.220

[2] Sun, Y., Wang, X., & Tang, X. (2014). DeepID: Face verification with deep neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1701–1708. https://doi.org/10.1109/CVPR.2014.220

[3] Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 815–823. https://doi.org/10.1109/CVPR.2015.7298682

[4] Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). ArcFace: Additive angular margin loss for deep face recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4690–4699. https://doi.org/10.1109/CVPR.2019.00482

[5] DeepInsight. (2025, August 29). InsightFace. GitHub. https://github.com/deepinsight/insightface

[6] Chen, S., Chen, C., Zhang, X., & Huang, K. (2018). MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices. In Z. Sun & C. Shan (Eds.), Biometric Recognition (CCBR 2018). Lecture Notes in Computer Science (Vol. 10996, pp. 428–438). Springer. https://doi.org/10.1007/978-3-030-02257-3_46

[7] Zhang, Q., Zhang, X., & Wang, Y. (2020). ShuffleFaceNet: A lightweight face recognition network for mobile devices. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 1–12. https://doi.org/10.1007/978-3-030-58536-5_29

[8] Huang, T., Li, X., & Zhang, Y. (2021). Differential privacy in face recognition: A survey. arXiv. https://arxiv.org/abs/2105.12182

[9] Peng, X., Li, Y., & Zhao, Z. (2022). Unsupervised domain adaptation for face recognition: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(12), 9649–9667. https://doi.org/10.1109/TPAMI.2021.3054824

[10] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT) (pp. 77–91). ACM. https://doi.org/10.1145/3287560.3287572