AI Engineering
https://www.sciltp.com/journals/aieng

Review

A Narrative Review on the Development and Evolution of Biometric Recognition

Kaihui Wu

College of Medicine and Biological Information Engineering, Northeastern University, Shengyang 110169, China; wukh@mails.neu.edu.cn

How To Cite: Wu, K. A Narrative Review on the Development and Evolution of Biometric Recognition. *AI Engineering* **2026**, *2*(2), 9. https://doi.org/10.53941/aieng.2026.100009

Received: 24 January 2026
Revised: 19 June 2026
Accepted: 24 June 2026
Published: 8 July 2026

Abstract: Biometric recognition based on physiological and behavioral traits has become a core supporting technology in the field of information security. This article provides a structured narrative review that systematically combs the development history of biometric recognition around six typical unimodal modules, namely iris, fingerprint, face, finger vein, ECG, and gait, as well as the technical progress of multimodal fusion. On this basis, this review sorts out representative research to show key technical breakthroughs, performance promotion and module expansion achieved in recent years. Meanwhile, typical challenges faced by the current field are analyzed, and feasible development trends in the future are discussed.

Keywords: biometric recognition; convolutional neural networks; machine learning; deep learning; iris recognition; facial recognition; fingerprint identification

1. Introduction

Driven by growing demands for secure, convenient, and efficient identity verification in smart environments, financial services, and public security scenarios, biometric recognition has gradually evolved from traditional manual verification toward intelligent automated systems. Different from password-centric authentication mechanisms, biometric technologies identify individuals according to innate physiological or behavioral characteristics, such as iris texture, fingerprint ridges, and walking patterns, that are hard to counterfeit, replicate, or lose [1]. Over the past decade, the field has undergone a major shift from handcrafted feature extraction to data-driven deep learning representation, while single-modal systems have increasingly evolved into multimodal fusion architectures to achieve stronger stability and reliability [2].

Single-modal biometric systems perform identity authentication and verification based on only one biological trait, such as fingerprint or voiceprint. Such systems feature mature pipelines and simple deployment, yet their performance is easily affected by environmental interference and individual physiological differences, resulting in limited accuracy in unconstrained application scenarios. Multimodal fusion integrates two or more biometric cues or multi-source data of the same trait, supporting information complementarity and comprehensive decision-making at the data, feature, and decision levels. This strategy effectively alleviates the inherent defects and performance bottlenecks of single-modal biometric systems.

This survey focuses on two core aspects: the development of key unimodal biometric modules and the evolution of multimodal fusion techniques. We select representative studies covering iris, fingerprint, face, finger vein, ECG, and gait recognition (a critical dynamic biometric trait), analyzing their technical routes, performance breakthroughs, and practical applications. The references include recent works on deep learning-based feature extraction [3], multimodal fusion frameworks [4], anti-spoofing strategies [5], and lightweight model design, providing a comprehensive overview of the field's development landscape.

2. Development of Unimodal Biometric Recognition Modules

According to literature research, unimodal biometric recognition is a technology that accomplishes identity authentication and verification relying solely on a single biometric trait. With a mature framework, it is a research hotspot in the interdisciplinary field of computer vision and bioinformatics. This section focuses on the key technological breakthroughs in iris recognition, fingerprint recognition, face recognition, gait recognition, finger vein recognition and ECG recognition.

2.1. Iris Recognition

Iris recognition remains one of the most reliable biometric solutions due to its strong uniqueness and long-term stability, especially for applications that demand high security standards. The theoretical and algorithmic basis of modern iris recognition-was formally established by Daugman through the rubber sheet model, 2D Gabor wavelet encoding, and statistical independence verification. These core components became the-basis of nearly all iris recognition systems developed in later years. Early implementations followed this handcrafted design using Gabor filters and wavelet transformation, yet often struggled with noise interference and partial occlusions in real-world imaging conditions [6,7]. Deep learning has since reshaped feature extraction in iris recognition: unlike conventional methods that require strict segmentation and normalization, learning-based frameworks such as IrisConvNet combine these steps into a unified end-to-end structure, delivering strong verification performance on the IITD database [3]. More recently, Vision Transformers enhanced with keypoint-dependent relative position encoding (KP-RPE) have further improved stability when dealing with misaligned or low-quality iris samples.

Several innovative CNN architectures have further enhanced the precision and generalization ability of iris recognition. ComplexIrisNet [8] is a fully complex-valued CNN that models iris texture in the complex domain, preserving both amplitude and phase information to overcome the drawbacks of real-valued networks in capturing subtle patterns. This architecture connects mathematically with classical Gabor wavelets and integrates traditional coding theory with data-driven feature learning, showing stronger cross-dataset generalization on noisy iris images. DualSANet [9] is a dual spatial attention network built on ResNet-18, equipped with a dual-branch decoder and Spatial Attention Feature Fusion Modules (SAFFM) to suppress interference from eyelids, eyelashes, and uneven illumination, thus improving occlusion robustness. On the CASIA-Iris-Thousand dataset, DualSANet reduces the equal error rate (EER) by 3.2% compared with baseline models. MS-DGR [10]: A multi-scale dynamic graph representation method that converts iris features into structured graphs at shallow, middle, and deep levels. It uses graph triplet loss to enhance robustness to partial occlusions, outperforming traditional CNNs by 5.7% in recognition accuracy on occluded iris samples.

IrisConvNet, ComplexIrisNet and DualSANet improved robustness to eyelid occlusion, uneven illumination and poor image quality, although state-of-the-art methods still fail on long-range, non-cooperative and motion-blurred iris images. In real-world scenarios, the equal error rate (EER) is usually much higher than the ideal laboratory values reported under constrained imaging conditions [11,12].

2.2. Fingerprint Recognition

As one of the most widely used biometric traits, fingerprint recognition has evolved from minutiae-based matching to deep learning-driven end-to-end systems: Early methods focused on extracting minutiae points (ridge endings, bifurcations) and template matching [13] perform poorly on blurred, stained or worn fingerprints, while deep learning models (e.g., CNN-based FVC2004-winning architectures) automatically learn discriminative features from raw images. Recent works like DeepPoreID utilize deep convolutional neural network (DCNN) for pore-level feature extraction, achieving EER of 0.16 [14].

Contemporary deep learning architectures have further strengthened the accuracy and stability of fingerprint recognition by addressing issues related to weak texture, noise, and sample variability. The encoder-residual-decoder structure [15] is a specialized CNN for phase extraction in non-contact 3D fingerprint imaging. It uses 7 × 7 convolution kernels and a sine-constrained loss function to reduce phase discontinuities, thereby retaining clearer ridge details. This design achieves an equal error rate (EER) of 1.8% on the FVC2004 DB1 dataset and performs better than conventional phase unwrapping techniques.

The SAMAF network [16] is an independent multi-attention architecture built for detecting mixed or duplicated fingerprints. It combines multi-instance attention (MIA) and multi-attention fusion (MAF) modules to enhance the detection of synthetic or altered fingerprint traces. The MIA component integrates spatial and channel attention, while MAF fuses multi-scale information, yielding a detection accuracy of 98.6% on datasets containing both real and synthetic samples.

As another advanced solution, IFViT [17] provides an interpretable fixed-length fingerprint matching framework based on twin Vision Transformers and random sample consensus (RANSAC) optimization. It captures long-range dependencies through self-attention and cross-sample comparison, bringing the EER on FVC2004 down to 1.87%, which is 45% lower than DeepPrint. The model also supports visual interpretation via feature heatmaps that reflect ridge-based decision processes.

Practical deployment still faces challenges posed by low-quality fingerprints caused by abrasion, contamination, or skin conditions, as well as spoofing attacks. These issues are commonly relieved through preprocessing operations such as TopHat filtering and density-based spatial clustering of applications with noise (DBSCAN) clustering [4], together with CNN-based liveness detection for anti-spoofing protection [18]. Data augmentation including rotation and reflection is also widely adopted to improve the generalization ability of recognition models [19].

2.3. Face Recognition

Face recognition is the most widely applied biometric technology, benefiting from the ubiquity of image capture devices. Its development reflects the impact of deep learning, particularly in the evolution of verification models: Traditional methods, like local binary pattern and histogram of oriented gradients (LBP, HOG), struggled with pose, illumination, and occlusion variations, while deep learning models (e.g., visual geometry group 16-layer network VGG-16, ArcFace) achieve state-of-the-art performance on labeled faces in the wild (LFW) and CelebA datasets [20]. Furthermore, recent advances include keypoint-dependent relative position encoding (KP-RPE) to handle misalignment [7] and quality-guided feature fusion to improve robustness [5].

Several specialized frameworks have extended facial recognition toward related understanding tasks. The finger vein enhancement and super-resolution (FESR) framework [18] constructs a joint learning pipeline that combines facial expression synthesis generative adversarial network (FESGAN) and identity recognition. FESGAN produces expression-rich images with consistent identity information, and the recognition network optimizes feature distribution using triplet loss on original, real, and synthetic samples. This design boosts small-sample facial expression recognition accuracy by 12.3% on the CK+ dataset. deep multi-task multi-label convolutional neural network (DMM-CNN) [19] represents a typical multi-task learning structure that unifies 72-point facial landmark detection and attribute classification. It applies 1-level and 3-level Spatial Pyramid Pooling (SPP) to extract features for objective and subjective attributes separately, and employs an adaptive weight assignment strategy to emphasize hard-to-classify attributes. On the CelebA dataset, the model reaches a mean classification accuracy of 91.7%.

In addition to the above architectures, partial fully connected strategy (Partial FC) [21]: A sparse update strategy for large-scale face recognition. It randomly samples partial negative class centers during training, reducing computational complexity from O(N) to O(K) (K < It supports 10-million-scale identity training on a single server, with 30% faster training speed than traditional full-connected layers. Meanwhile, AdaFace [22]: A quality-adaptive margin loss function that dynamically adjusts angular margin based on feature norm (a proxy for image quality). High-norm samples use larger margins to enhance separability, while low-norm samples use smaller margins to avoid overfitting. It reduces EER by 27% on low-quality face datasets compared to ArcFace.

Anti-Spoofing Progress: Face forgery attacks (deepfakes, 3D masks) have driven the development of anti-spoofing techniques. Deep learning-based methods focus on texture analysis (high-frequency features) and supervised contrastive learning to address data imbalance, achieving average classification error rate (ACER) as low as 0.06 [23]. However, face recognition is extremely vulnerable to presentation attacks including photos, videos, 3D masks, and deepfakes. Even advanced anti-spoofing methods cannot fully eliminate security risks in open, unconstrained scenarios.

The extensive deployment of facial recognition has triggered widespread disputes over privacy security and legal compliance across the globe. Facial data, as a kind of irreplaceable sensitive biometric information, is prone to unauthorized collection, covert tracking, data breaches and abusive application once lacking strict supervision, which may seriously infringe personal rights and interests. Many countries and regions have issued targeted regulatory policies to restrict improper usage scenarios, such as unrestricted public surveillance, cross-border data transmission and commercial utilization without sufficient consent. In addition, problems such as algorithmic discrimination, insufficient security protection and incomplete consent mechanism further intensify social worries and promote the improvement of relevant legal systems [24]. How to coordinate technological progress, public security needs and fundamental privacy rights has become a key issue to promote the standardized and healthy development of facial recognition systems.

At present, deep learning serves as the mainstream scheme for face recognition, state-of-the-art algorithms achieve outstanding performance under constrained datasets, yet severe occlusion, low-quality inputs and presentation attacks remain intractable bottlenecks that cannot be fully eliminated. Future research needs to simultaneously balance recognition accuracy, anti-spoofing capability and privacy compliance requirements.

2.4. Gait Recognition

Gait recognition-recognition characterizes individuals through distinctive walking patterns, and has drawn extensive attention owing to its non-contact nature and capability for long-distance identification. CNN-based techniques have greatly promoted the development of gait recognition, especially in spatiotemporal feature modeling and cross-view adaptability. Two milestone methods have established the mainstream paradigm of modern gait recognition. GaitSet [25] treats gait sequences as an unordered frame set and realizes strong cross-view generalization ability through set-based feature aggregation, which has become a typical reference for subsequent studies. GaitGL [26] further strengthens discriminative capacity by fusing global appearance information, local component features, and temporal aggregation, thus improving recognition stability under changes in clothing and viewing angle.

In addition to deep learning-based paradigms, several early non-intrusive systems have also provided valuable attempts for gait recognition research. A typical instance is BRITTANY (Biometric RecognITion Through gAit aNalYsis) [27], a non-contact gait identification system developed by the University of León, Spain. This system adopts 2D laser scanning technology and machine learning algorithms, and uses the PeTra (People Tracking) module to process raw laser data, generate spatial distribution maps, extract human contours, and track moving objects continuously. Through an integrated pipeline of target tracking, walking pattern analysis and model classification, BRITTANY achieves indoor identity authentication and illegal intrusion detection, with an identification accuracy of 99%.

On the basis of these early attempts, event-driven and adaptive spatiotemporal methods have further broadened the application potential of gait recognition. EV-Gait-3DGraph and EV-Gait-IMG [28] offer efficient event-stream representation schemes for low-power consumption scenarios, while 3D Local CNN [29] captures fine-grained motion features through adaptive spatiotemporal convolutions. Furthermore, attribute-augmented spatial-temporal graph convolutional network (AST-GCN) [30] extends gait analysis to emotion recognition by constructing enhanced spatial-temporal graph correlations, proving the flexible extensibility of gait-based biometric techniques.

Despite these advances, the core bottlenecks remain: First, strong appearance interference, where clothing, carried items and body shape changes cover the core gait features, leading to significant accuracy attenuation in cross-clothing and cross-pose recognition; second, difficult cross-view modeling, random collection perspectives in actual scenes cause the offset of gait feature distribution, and the generalization of traditional models is insufficient; third, spatiotemporal feature coupling, it is difficult to decouple spatial limb shape and temporal movement laws while balancing modeling efficiency and feature expression ability. In practical application landing, it also faces prominent difficulties: high hardware transformation costs, which require high-precision visual acquisition equipment such as event cameras and depth cameras, and low-quality gait data collected by ordinary surveillance cameras cannot meet recognition needs; strict scene constraints, it is only suitable for non-contact long-distance recognition, and dense crowds and severely occluded scenes will cause gait sequence breakage and failure to extract effective features [31].

Current gait recognition mainly adopts set-based modeling and spatial-temporal graph convolution as mainstream methods, and its unique advantage lies in long-distance non-cooperative identification. Occlusion from clothing, cross-view variation and high hardware cost are key bottlenecks restricting practical deployment. Future research of gait recognition will mainly focus on relieving computational pressure brought by complex models such as 3D CNNs and realizing efficient reasoning on edge devices. The development directions include lightweight network design, efficient cross-view feature alignment, and adaptive modeling in complex scenes [25].

2.5. Finger Vein Recognition

Finger vein recognition, leveraging subcutaneous vein patterns, offers high anti-spoofing capability due to its internal biological nature [32]. Its development has accelerated with deep learning: especially in the field of feature learning: Early methods relied on vein pattern enhancement and template matching, while deep learning models (e.g., CNN, GAN) extract discriminative features from near-infrared images.

In addition to feature learning progress, Siamese CNN architecture has become a classic metric learning paradigm for finger vein verification, which addresses the weak discriminability of shallow handcrafted vein features via paired sample training and modified contrastive loss (MC loss) [33]. Tang et al. proposed a two-stage training pipeline combining data augmentation, pre-trained CNN and knowledge distillation to relieve the shortage of finger vein training samples, and further compressed the network into a tiny lightweight model suitable for embedded mobile terminals, reaching an EER of 0.08% on MMCBNU_6000 dataset [33]. Key public datasets widely adopted in existing research include SDUMLA-HMT, MMCBNU_6000, and University of Twente Finger Vein Pattern (UTFVP), with lightweight models designed for mobile applications [34]. Another prominent advantage of finger vein recognition lies in its fusion potential. Finger vein is often fused with fingerprint or face to enhance reliability.

Finger vein recognition has prominent core technical bottlenecks: First, it is sensitive to image quality, near-infrared collection is easily affected by dry fingers, skin folds and collection angles, leading to blurred vein texture and missing features; second, the difficulty of feature extraction is high, finger vein texture is a shallow subcutaneous feature with low texture contrast, making it difficult to distinguish from background noise, and traditional handcrafted features are difficult to capture effective discriminant information; third, insufficient sample size, the vein texture of different groups such as the elderly and children varies greatly, and the sample size of public datasets such as MMCBNU_6000 is far less than that of fingerprint and face datasets, leading to easy overfitting of models. In practical application landing, it faces three major difficulties: low standardization of collection equipment, large differences in wavelength and resolution of near-infrared collection equipment from different manufacturers, leading to incompatibility of cross-device vein features and poor model generalization; poor user experience, collection requires fingers to be close to the collection window, and strict requirements on collection angle and finger placement posture, with lower user operation convenience compared with fingerprint and face recognition; high difficulty in fusion application, although fusion with fingerprint and face can improve system reliability, multi-modal feature collection requires the cooperation of multiple devices, increasing device volume and deployment costs, and it is difficult to adapt to miniaturized terminals such as smart bracelets.

2.6. ECG (Electrocardiogram) Recognition

ECG (Electrocardiogram) recognition, based on unique cardiac electrical activity, excels in liveness detection, making it ideal for anti-spoofing scenarios. Its development focuses on signal processing and deep learning integration, with remarkable progress in technical evolution: Traditional methods extract features from QRS complexes (ventricular depolarization waveforms in ECG signals) and T-waves, while deep learning models transform 1D ECG signals into 2D images for CNN-based feature extraction [35]. The 9-layer CNN architecture can automatically learn discriminative waveform features without manual feature engineering, and it maintains stable classification accuracy even on unfiltered noisy ECG data. After data balance augmentation, the model reaches 94.03% classification accuracy on denoised ECG samples and outperforms classic machine learning algorithms including LDA and FL on the MIT-BIH dataset [35]. Beyond technical evolution, ECG is often fused with fingerprint to enhance security.

Acquisition poses a major challenge and acts as a key limiting factor in ECG biometric systems. Typical ECG recording requires fixed electrodes and stable physical conditions, making it hard to apply in daily environments. Non-contact designs still suffer from poor stability and are easily disrupted by body motion and external noise. Even mild changes in physical or mental state can distort cardiac signals, making consistent feature extraction difficult. Such problems in signal collection directly restrict the practical performance of ECG identification [36].

ECG recognition has three core technical bottlenecks: First, serious signal noise interference, ECG signals are easily affected by human movement, environmental electromagnetism and poor contact of collection equipment, and noise covers core features such as QRS waves and T waves; second, significant individual differences, the heart rate and ECG waveform of different groups vary greatly, and the ECG features of the same person change under different physiological states such as tension and post-exercise, requiring high model robustness; third, single signal dimension, ECG is a 1D time-series signal with limited feature expression ability, and how to transform time-series signals into highly discriminant feature representations is a core difficulty in deep learning modeling. In practical application landing, it also faces prominent difficulties: poor collection convenience, traditional ECG collection requires professional electrode sheets and equipment, non-contact ECG collection technology is not yet mature, making it difficult to achieve rapid collection for daily identity authentication; strict physiological state constraints, ECG features are greatly affected by human physiological states, and signals collected after exercise and emotional fluctuations cannot be effectively identified, resulting in limited applicable scenarios (such as only suitable for static identity authentication); difficult edge-end processing, real-time noise reduction and feature extraction of ECG signals require special algorithms, and ordinary mobile terminals lack corresponding hardware acceleration modules, leading to low real-time processing efficiency.

With the continuous advancement of related technologies, ECG recognition is widely used in healthcare and mobile authentication, and to better adapt to practical needs, recent works focusing on noise reduction and real-time processing [35].

To intuitively compare the models, datasets, evaluation metrics, inherent strengths and limitations of the six mainstream single-modal biometric recognition technologies mentioned above, a comprehensive comparative summary table is presented in Table 1.

Table 1. Performance Comparison of Six Typical Single-modal Biometric Recognition Modalities.

Modality	Representative Model	Dataset	Key Metric	Main Advantages	Limitations
Iris	IrisConvNet, ComplexIrisNet, DualSANet, MS-DGR	IITD, CASIA-Iris-Thousand	Accuracy, EER	Extremely high uniqueness and long-term biological stability Complex-valued CNN and spatial attention effectively suppress occlusion and illumination noise	Poor performance for long-distance, motion-blurred, non-cooperative iris images High equipment cost and strict acquisition constraints
Fingerprint	DeepPoreID, Encoder-residual-decoder, SAMAF, IFViT	FVC2004 DB1, Multi-source mixed fingerprint dataset	EER, Detection Accuracy	Mature industrial deployment, low hardware cost Pore-level feature extraction and ViT realize interpretable matching	Susceptible to abrasion, stains, dry skin and physical damage Easy to be attacked by artificial fingerprint molds
Face	ArcFace, Partial FC, AdaFace, FESR, DMM-CNN	LFW, CelebA, CK+	Accuracy, EER, ACER	Wide acquisition equipment coverage, non-contact convenient collection Multi-task joint learning supports expression recognition and attribute classification	Vulnerable to photo, video, 3D mask and deepfake presentation attacks Severe performance degradation under heavy occlusion, large pose and poor lighting
Gait	GaitSet, GaitGL, 3D Local CNN, AST-GCN, EV-Gait series	CASIA-B, Laser gait dataset	Identification Accuracy	Long-distance non-cooperative recognition without user cooperation Event-stream lightweight models adapt to low-power edge devices	Recognition accuracy drops drastically with clothing changes and object occlusion High requirements for depth camera/event camera hardware
Finger Vein	FVRAS-Net	SDUMLA-HMT, MMCBNU_6000, UTFVP	Recognition Rate, EER	Subcutaneous biological feature with strong inherent anti-spoofing ability Lightweight models support mobile terminal deployment	Extremely sensitive to finger dryness, skin folds and collection angles Small public dataset scale, easy model overfitting
ECG	ECG-CNN, 1D-to-2D transformation convolutional network,	MIT-BIH arrhythmia dataset	AUC, Classification Accuracy, EER	Unique cardiac electrical signal, absolute advantage in liveness detection Can be combined with medical health monitoring scenarios	Signals disturbed by exercise, mood and electromagnetic noise Traditional collection relies on fixed electrodes with poor convenience

3. Evolution of Multimodal Biometric Fusion

3.1. Development of Fusion Strategies

Multimodal biometric fusion emerged to overcome the limitations of unimodal systems (e.g., face recognition's sensitivity to lighting, fingerprint's vulnerability to damage). Fusion strategies are typically categorized into three levels, each with distinct advantages, disadvantages, and application scenarios:

Score-level fusion combines matching results from different modalities through simple rules such as summation, maximization, and product calculation. It is lightweight and easy to implement in most frameworks, but cannot retain complete feature details, resulting in relatively weak robustness. It is often used in edge devices with limited resources and low-complexity application scenarios [8].

Feature-level fusion integrates high-dimensional features directly before classification. This strategy keeps the most effective feature information and achieves better accuracy, but brings higher computation and potential dimensional problems. Therefore, it is more suitable for server-side scenarios that focus on precision rather than real-time speed [37].

Decision-Level Fusion integrates the final classification results of each modality via logical rules or meta-learners. It is flexible, modular, and robust to missing modalities, but has lower accuracy than feature-level fusion. It is suitable for heterogeneous systems, multi-sensor platforms, and scenarios with random modality missing [19].

3.2. Key Technical Advancements

Recent studies have advanced multimodal fusion in three aspects: The first aspect is deep learning-driven fusion, where models like CNN and Transformer learn cross-modal correlations automatically. The FarSight system fuses face, gait, and body shape via quality-guided Mixture-of-Experts (QME), adapting to varying modality quality [7]. The second aspect focuses on missing modality processing, aiming to solve the problem of performance attenuation caused by the loss of partial modal features due to collection failure and occlusion in actual scenarios. The core technical paths are divided into feature completion and modal decoupling: feature completion generates the feature representation of the missing modality according to the existing modal features through generative models such as GAN and variational autoencoder (VAE), and then performs multimodal fusion, which is suitable for scenarios with a low proportion of missing modalities; modal decoupling decomposes multi-modal features into shared features (identity features common to each modality) and specific features (unique features of each modality), and only uses shared features during fusion. Even if some modalities are missing, the shared features can still ensure recognition performance, which is suitable for scenarios with random modal missing. Representative studies include SSFD-Net [9].

The third aspect is cross-sensor fusion, which addresses the problem of inconsistent feature distribution of multi-modal data collected by different types and manufacturers of sensors (such as visual sensors collecting face/gait, near-infrared sensors collecting finger veins, and bioelectric sensors collecting ECG). The core technical paths are sensor domain adaptation and unified feature space construction: sensor domain adaptation eliminates the domain gap of data collected by different sensors through methods such as adversarial learning and domain normalization, making multi-modal features have similar distributions in the same feature space; unified feature space construction designs a general feature extraction backbone network, performs unified feature encoding on the original data of different sensors, and then integrates the encoded features through a fusion module to realize effective fusion of cross-sensor features.

In addition, the efficient fusion architectures, where lightweight frameworks (e.g., Finger-Vein Recognition and AntiSpoofing Network FVRAS-Net) integrate recognition and anti-spoofing into a single model, reducing computational cost for edge devices [38], it has become a key research direction for multimodal fusion to adapt to the low computing power and low storage deployment requirements of edge devices such as mobile terminals, IoT devices and security cameras. Its core technical paths focus on model compression and lightweight fusion layer: model compression compresses the single-modal feature extraction network by means of depthwise separable convolution, knowledge distillation and sparsification to reduce the computational cost of single-modal feature extraction; lightweight fusion layer simplifies traditional complex fusion modules (such as fully connected layer fusion, attention fusion), and adopts lightweight fusion methods such as feature concatenation + lightweight convolution and weighted summation + channel attention to reduce the computational complexity of the fusion layer on the premise of ensuring fusion performance.

Deep learning-driven cross-modal modeling, missing modality compensation and lightweight fusion architectures are popular research hotspots, which can effectively compensate the inherent defects of unimodal biometrics. Nevertheless, cross-sensor domain shift and high deployment cost of multi-device systems still hinder large-scale practical applications, and future studies will focus on end-to-end lightweight fusion and privacy-preserving fusion frameworks.

4. Core Challenges and Technical Breakthroughs

4.1. Spoofing Attacks and Anti-Spoofing Techniques

Biometric systems face serious security risks from spoofing attacks, including deepfake media, 3D facial masks, and artificial fingerprint replicas. Recent studies have proposed several effective detection methods.

Texture-aware detection models target high-frequency details in facial samples, such as subtle surface roughness and local color variations that fake samples struggle to reproduce. This kind of detection has achieved an ACER of 0.06 in relevant experiments [23]. Liveness detection provides a further layer of protection. ECG-based verification can distinguish living individuals from fake samples, as artificial traits cannot generate realistic cardiac signals [19]. Deep learning approaches also improve anti-spoofing performance. Supervised contrastive learning is widely used to spot GAN-manufactured fake samples, easing the problem of imbalanced real and fake training data [23].

4.2. Computational Complexity and Lightweight Design

A critical common challenge across all unimodal biometric recognition modules is high computational complexity, which severely hinders the real-time deployment of models on resource-constrained edge devices (e.g., mobile phones, IoT sensors, embedded security terminals). This challenge is the core motivation for the development of lightweight design techniques, which have formed a systematic optimization system through continuous innovation and are further extended to multimodal fusion scenarios to construct a complete single-modal lightweight—lightweight multimodal fusion technical chain for the whole biometric recognition system. Recent representative lightweight innovation strategies are as follows:

Sparse Attention Mechanisms [39]: Dynamically activate key spatiotemporal nodes (e.g., gait cycle keyframes, facial landmarks) for feature extraction, reducing computational cost by 40–60% without significant accuracy loss, and is particularly suitable for dynamic biometric recognition modules (gait, ECG) with spatiotemporal feature characteristics.

Knowledge Distillation [40]: Compress large pre-trained models into lightweight versions by transferring knowledge from teacher models to student models. A typical example is MobileFaceNets, which achieves 99.2% accuracy on LFW with only 4.5 M parameters (1/10 of VGG-16) and has become a mainstream lightweight framework for face recognition on mobile terminals.

Hybrid Architectures [41]: Combine CNN's local perceptual advantage with ViT's global feature modeling capability to balance model performance and computational complexity. MobileViT, for instance, achieves Swin Transformer-level recognition accuracy with 50% fewer FLOPs, and is widely applicable to various unimodal modules (iris, fingerprint, face) for edge device deployment.

The above lightweight design strategies provide universal optimization methods for all unimodal biometric recognition modules. On this basis, the strategies are extended to multimodal fusion systems to solve the computational complexity problem of multi-feature integration, realizing end-to-end lightweight optimization of biometric recognition from single-modal feature extraction to multi-modal fusion.

4.3. The Limitations

The limitations of biometric recognition are mainly reflected in ethics, physical occlusion, acquisition adaptability, security, and cost.

At the ethical level, biometric characteristics are lifelong and immutable. Large-scale collection and networking of such characteristics can easily trigger concerns about "Orwellian" surveillance (totalitarian monitoring), leading to potential privacy abuse, data leakage, power abuse, and eventually social exclusion [42]. For example, once biometric data is leaked or stolen, it cannot be reset like a password, which may result in identity theft and phishing attacks. Moreover, compromised biometric traits can be reused across multiple systems, posing lifelong privacy and security risks to individuals.

To date, a globally unified legislative protection system for biometric data has not yet been formed. There are significant divergences among countries in the authorization boundaries, scope of use, and cross-border transmission rules of data collection. The laws and regulations in some regions only provide principled provisions for biometric data, lacking detailed constraints on the entire process of collection, storage and use. This leads to excessive operational flexibility for enterprises and institutions, which is prone to social conflicts caused by data abuse. Meanwhile, the application of biometric recognition technology has the problem of a “digital divide”. Groups with weak operational capabilities of intelligent devices, such as the elderly and the disabled, are prone to face service exclusion due to the operational threshold of biometric collection and recognition, which exacerbates social inequality and violates the sociological principle of technological inclusiveness.

Meanwhile, some systems have stringent requirements for environmental conditions and user cooperation, resulting in poor performance in real-world, uncontrolled and complex scenarios. For instance, the performance of ECG recognition is severely restricted during the signal acquisition stage [36]. Acquisition-related problems lead to feature instability and high false detection rates, creating a significant performance gap between laboratory-controlled environments and real-world applications. Even highly mature technologies such as face recognition suffer from sharp accuracy degradation when users are uncooperative, and simple physical occlusion is sufficient to substantially reduce recognition performance.

The above limitations are not irreparable. With technological iteration (such as more advanced liveness detection, algorithm optimization, and fusion of multi-modal biometrics) and the improvement of relevant laws and regulations, some problems can be alleviated but cannot be fundamentally eliminated. In practical applications, it is usually necessary to combine multiple authentication methods (such as biometrics + password) to balance security, accuracy, and usability.

4.4. Performance Optimization

Key performance indicators including identification accuracy, equal error rate (EER), and area under curve (AUC) have been steadily improved in recent biometric studies. Prominent progress has been made in unimodal recognition tasks covering iris, fingerprint, ECG, and gait. For multimodal recognition, advances are concentrated in sequential fusion, parallel fusion, and long-range identity authentication. The performance of several representative deep learning-based models is summarized in Table 2.

Table 2. Deep Learning Models in Biometric Recognition.

Ref.	Modality	Task Type	Foundation Model	Dataset	Key Metric	Core Innovation	Limitation
[3]	Iris	Verification	IrisConvNet	IITD	Acc = 99.2%	End-to-end segmentation & feature extraction	Poor on low-quality & non-cooperative samples
[4]	Face + Iris + Vein	Verification	FRCNN	SDUMLA-HMT	Acc = 99.53%	Feature-level multimodal fusion	Poor cross-sensor generalization
[8]	Iris	Verification	ComplexIrisNet	CASIA-Iris-Thousand	EER = 0.12%	Complex-valued feature encoding	High computational cost
[17]	Fingerprint	Verification	IFViT (ViT)	FVC2004	EER = 1.87%	Interpretable fixed-length representation	Weak on severely blurred fingerprints
[22]	Face	Verification	AdaFace	LFW	Acc = 99.83%	Quality-adaptive angular margin	Degraded under heavy occlusion
[26]	Face + Gait	Identification	FarSight (ViT)	Long-range	TAR@0.1%FAR = 83.1%	Quality-guided multimodal fusion	High hardware dependence
[29]	Gait	Identification	3D Local CNN	CASIA-B	Acc = 94.1%	Multi-temporal-scale spatiotemporal fusion	Sensitive to clothing & view changes
[33]	Finger Vein	Verification	Siamese CNN	MMCBNU_6000	EER = 0.08%	Siamese paired training + modified contrastive loss	Vulnerable to finger dryness, uneven near-infrared illumination

5. Future Research Directions

5.1. Technical Trends

Lightweight Models: Developing high-efficiency and compact network architectures tailored for resource-constrained edge devices including mobile phones, IoT sensors, and embedded security terminals. The research focuses on optimizing model structure via techniques such as depthwise separable convolution, knowledge distillation, and sparse attention mechanism, minimizing computational overhead and storage footprint while maintaining state-of-the-art recognition performance, thus enabling real-time and low-power biometric authentication in edge computing scenarios [41].

Privacy-Preserving Biometrics: Fusing core cryptography technologies to build a dual-protection privacy preservation system of "technology + regulation". Key research directions include developing distributed training schemes combining federated learning with gradient encryption, lightweight homomorphic encryption inference models, and high-discriminative revocable biometric transformation algorithms. Meanwhile, exploring the engineering application of zero-knowledge proof and trusted execution environments in biometric authentication to achieve the privacy protection goal of "data available but invisible" [2].

Interdisciplinary fusion optimization: Combining bioinformatics achievements to construct biometric sample libraries for different populations (children, the elderly, different races, and the disabled), and developing adaptive extraction and matching algorithms for specific features to improve technological universality; collaborating with researchers in law and sociology to participate in the formulation and improvement of laws and regulations on biometric data, clarify the ethical boundaries of technological application, and design technology inclusiveness schemes to lower the use threshold for the elderly, the disabled and other groups, alleviating social conflicts caused by technological application.

5.2. Application Expansions

Long-Range and Unconstrained Recognition: Extending biometric recognition technology to large-scale and unconstrained surveillance scenarios such as public security and smart city management, including UAV-based long-distance biometric recognition at the kilometer level and multi-target simultaneous recognition in dense crowd scenes [26]. The research focuses on integrating multi-modal features such as face, gait and body shape, and developing high-sensitivity feature extraction and matching algorithms adapted to long-distance, low-resolution and complex background conditions, to realize non-contact, non-cooperative and real-time identity recognition in unconstrained outdoor environments.

Medical Biometrics: Applying biometric recognition technology to the medical health field, using radiological images (chest X-rays), medical imaging data and physiological signals (ECG, EEG) for patient re-identification in healthcare information systems [10]. On the one hand, it realizes accurate matching of patient identity and medical records, solving the problems of medical record confusion and misdiagnosis caused by identity information errors; on the other hand, it combines biometric recognition with disease diagnosis, developing integrated systems of identity authentication and clinical disease auxiliary diagnosis, and providing personalized medical services based on the correlation between biometric features and physiological health status.

Multimodal IoT Integration: Deeply combining biometric recognition technology with IoT devices and systems, building a biometric-based security authentication system for the IoT ecosystem covering smart home, smart city, and industrial control. Realizing identity authentication and access control of IoT terminal devices, data transmission nodes and system operation personnel through multi-modal biometric features, solving the security risks such as illegal access and data tampering in the IoT system, and improving the overall security and reliability of the IoT ecosystem [19].

Interdisciplinary application innovation: Combining medicine and bioinformatics to develop an integrated system of disease auxiliary diagnosis and identity authentication based on physiological signals such as ECG and EEG, realizing the dual functions of "identity authentication + health monitoring" in medical scenarios; integrating cryptography and IoT technology to build a privacy-preserving biometric-based identity authentication system for IoT devices, solving the terminal security problem of IoT devices; fusing law and forensic technology to improve the judicial identification standards for biometric recognition evidence, and promoting the standardized application of technology in judicial forensics [41].

6. Conclusions

Biometric recognition has evolved gradually from manual feature design to deep learning–based single-modal systems, and further to stable multimodal fusion frameworks. This paper reviews the technical development of six typical unimodal biometrics: iris, fingerprint, face, finger vein, gait, and ECG. We also summarize multimodal fusion schemes at the score, feature, and decision levels.

Three main observations can be summarized: Deep learning has become the mainstream method in biometric recognition, greatly improving accuracy, anti-noise ability, and end-to-end learning. However, even state-of-the-art models still perform poorly in unconstrained environments, especially under non-cooperative conditions and low-quality inputs.

Multimodal fusion makes up for the natural flaws of single-modal systems. The choice between score-level, feature-level, and decision-level fusion depends mainly on accuracy demands, computing resources, and differences in sensing devices.

Real-world application is still limited by three key problems: high model complexity, spoofing attacks, and irreversible privacy risks. These issues require combined solutions such as lightweight networks, anti-spoofing mechanisms, and privacy protection techniques.

In general, biometric recognition is moving toward lightweight deployment, privacy-oriented design, long-distance non-cooperative sensing, and cross-disciplinary integration. Although great progress has been achieved in laboratory settings, narrowing the performance gap between controlled datasets and real scenes remains the central goal of future work. With proper algorithms, ethical rules, and proper regulations, biometric technology will continue to provide safe, efficient, and convenient identity authentication for smart cities, the Internet of Things, public security, and medical systems.

Funding

This research received no external funding

Data Availability Statement

This manuscript is a review, and no novel experimental data, raw datasets or samples were generated during the preparation of this paper. All information and evidence discussed in this work derive from previously published studies listed in the reference section. All cited publications are publicly accessible

Conflicts of Interest

The author declare no competing financial interests.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the author used ChatGPT to assist in searching for and screening relevant literature and used Deepseek and Kimi to translate text and polish sentences. After using this tool, the author reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

1. Jain, A.K.; Ross, A.; Prabhakar, S. An introduction to biometric recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2004**, *14*, 4–20. https://doi.org/10.1109/TCSVT.2003.818349.
2. Alay, N.; Al-Baity, H. Deep Learning Approach for Multimodal Biometric Recognition System Based on Fusion of Iris, Face, and Finger Vein Traits. *Sensors* **2020**, *20*, 5523.
3. Al-Waisy, A.S.; Qahwaji, R.; Ipson, S. A multi-biometric iris recognition system based on a deep learning approach. *Pattern Anal. Appl.* **2018**, *21*, 783–802.
4. Wang, Y.; Liu, Z. A Survey on Multimodal Biometrics. In *Advances in Automation and Robotics*; Lee, G., Eds.; Springer, Berlin, Heidelberg, 2011 https://doi.org/10.1007/978-3-642-25646-2_51.
5. Yu, Z.; Qin, Y.; Li, X.; et al. Multi-Modal Face Anti-Spoofing Based on Central Difference Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 2766–2774. https://doi.org/10.1109/CVPRW50498.2020.00333.
6. Daugman, J.G. High confidence visual recognition of persons by a test of statistical independence. *IEEE Trans. Pattern Anal. Mach. Intell.* **1993**, *15*, 1148–1161. https://doi.org/10.1109/34.244676.

7. Daugman, J. How iris recognition works. *IEEE Trans. Circuits Syst. Video Technol.* **2004**, *14*, 21–30. https://doi.org/10.1109/TCSVT.2003.818350.
8. Nguyen, K.; Fookes, C.; Sridharan, S. Complex-valued iris recognition network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 182–196.
9. Yang, K.; Xu, Z.; Fei, J. DualSANet: Dual spatial attention network for iris recognition. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Virtual Conference, 5–9 January 2021; pp. 888–896.
10. Ren, M.; Wang, Y.; Zhu, Y. Multiscale dynamic graph representation for biometric recognition with occlusions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 15120–15136.
11. Nguyen, K.; Proença, H.; Alonso-Fernandez, F. Deep Learning for Iris Recognition: A Survey. *ACM Comput. Surv.* **2024**, *56*, 1–35. https://doi.org/10.1145/3651306.
12. Boyd, A.; Fang, Z.; Czajka, A.; et al. Iris presentation attack detection: Where are we now? *Pattern Recognit. Lett.* **2020**, *138*, 483–489.
13. Jain, A.; Hong, L.; Bolle, R. On-line fingerprint verification. *IEEE Trans. Pattern Anal. Mach. Intell.* **1997**, *19*, 302–314. https://doi.org/10.1109/34.587996.
14. Lin, C.; Kumar, A. Contactless and partial 3D fingerprint recognition using multi-view deep representation. *Pattern Recognit.* **2018**, *83*, 314–327.
15. Wang, H.X.; Shan, L.B.; Pang, Q.L. Phase extraction method of fingerprint fringe pattern based on convolutional neural network. *Acta Photonica Sin.* **2022**, *51*, 372–385.
16. Bhattacharya, G.; Puhan, N.B. Stand-alone multi-attention fusion network for double-identity fingerprint detection. *IEEE Trans. Biom. Behav. Identity Sci.* **2022**, *4*, 596–602.
17. Qiu, Y.; Chen, H.; Dong, X. IFViT: Interpretable fixed-length representation for fingerprint matching via vision transformer. *IEEE Trans. Inf. Forensics Secur.* **2025**, *20*, 559–573.
18. Yan, Y.; Huang, Y.; Chen, S. Joint deep learning of facial expression synthesis and recognition. *IEEE Trans. Multimed.* **2020**, *22*, 2792–2807.
19. Mao, L.; Yan, Y.; Xue, J.H. Deep multi-task multi-label CNN for effective facial attribute classification. *IEEE Trans. Affect. Comput.* **2022**, *13*, 818–828.
20. Shu, Y.; Yan, Y.; Chen, S.; et al. Learning spatial-semantic relationship for facial attribute recognition with limited labeled data. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 11911–11920.
21. An, X.; Deng, J.; Guo, J.; et al. Killing two birds with one stone: Efficient and robust training of face recognition CNNs by partial FC. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 4032–4041.
22. Kim, M.; Jain, A.K.; Liu, X. AdaFace: Quality adaptive margin for face recognition. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 18729–18738.
23. Kim, H. Novel Deep Learning-Based Facial Forgery Detection for Effective Biometric Recognition. *Appl. Sci.* **2025**, *15*, 3613.
24. Robles, P.; Mallinson, D.J.; Best, E. Global perspectives on regulating facial recognition technology utilization for criminal justice arrests. *Glob. Public. Policy Gov.* **2025**, *5*, 186–204. https://doi.org/10.1007/s43508-025-00117-9.
25. Chao, H.; Wang, K.; He, Y.; et al. GaitSet: Cross-View Gait Recognition Through Utilizing Gait as a Deep Set. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 3467–3478. https://doi.org/10.1109/TPAMI.2021.3057879.
26. Lin, B.; Zhang, S.; Yu, X. Gait Recognition via Effective Global-Local Feature Representation and Local Temporal Aggregation. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021, pp. 14628–14636. https://doi.org/10.1109/ICCV48922.2021.01438.
27. Álvarez-Aparicio, C.; Guerrero-Higueras, M.; González-Santamarta, M. Biometric recognition through gait analysis. *Sci. Rep.* **2022**, *12*, 14530. https://doi.org/10.1038/s41598-022-18806-4.
28. Wang, Y.; Zhang, X.; Shen, Y. Event-stream representation for human gaits identification using deep neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 3436–3449.
29. Huang, Z.; Xue, D.; Shen, X.; et al. 3D local convolutional neural networks for gait recognition. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 14900–14909.
30. Chen, C.; Sun, X.; Tu, Z. AST-GCN: Augmented spatial temporal graph convolutional neural network for gait emotion recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 4581–4595.
31. Li, J. Rethinking Appearance-Based Deep Gait Recognition: Reviews, Analysis, and Insights from Gait Recognition Evolution. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 9777–9797. https://doi.org/10.1109/TNNLS.2025.3526815.

32. Kono, M.; Ueki, H.; Umemura, S.I. Near-infrared finger vein patterns for personal identification. *Appl. Opt.* **2002**, *41*, 7429–7436.
33. Tang, S.; Zhou, S.; Kang, W.; et al. Finger vein verification using a Siamese CNN. *IET Biom.* **2019**, *8*, 306–315. https://doi.org/10.1049/iet-bmt.2018.5245.
34. Sun, Z.; Feng, C.; Patras, I.; et al. LAFS: Landmark-based facial self-supervised learning for face recognition. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 1639–1649.
35. Acharya, U.R.; Oh, S.L.; Hagiwara, Y.; et al. A deep convolutional neural network model to classify heartbeats. *Comput. Biol. Med.* **2017**, *89*, 389–396. https://doi.org/10.1016/j.compbiomed.2017.08.022.
36. Pereira, T.M.C.; Conceição, R.C.; Sencadas, V.; et al. Biometric Recognition: A Systematic Review on Electrocardiogram Data Acquisition Methods. *Sensors* **2023**, *23*, 1507. https://doi.org/10.3390/s23031507.
37. Haghighat, M.; Abdel-Mottaleb, M.; Alhalabi, W. Discriminant correlation analysis for feature level fusion with application to multimodal biometrics. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 1866–1870. https://doi.org/10.1109/ICASSP.2016.7472000.
38. Yang, W.; Luo, W.; Kang, W.; et al. FVRAS-Net: An Embedded Finger-Vein Recognition and AntiSpoofing System Using a Unified CNN. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 8690–8701. https://doi.org/10.1109/TIM.2020.3001410.
39. Child, R.; Gray, S.; Radford, A.; et al. Generating long sequences with sparse transformers. *arXiv* **2019**, arXiv:1904.10509.
40. Chen, S.; Liu, Y.; Gao, X.; et al. MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices. In *Biometric Recognition*; Springer International Publishing: Cham, Switzerland, 2018; pp. 428–438.
41. Mehta, S.; Rastegari, M. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv* **2021**, arXiv:2110.02178.
42. Matulionyte, R.; Zalnieriute, M. Introduction: Facial Recognition in the Modern State. In *The Cambridge Handbook of Facial Recognition in the Modern State. Cambridge Law Handbooks*; Matulionyte, R., Zalnieriute, M., Eds.; University Press: Cambridge, UK, 2024; pp. 1–8.