Journal of Computer and Communications, 2026, 14(4), 103-118
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

A Review of Pedestrian Recognition Based on Millimeter-Wave Radar and Video Fusion

Hongtao Li, Qimeng Lu, Huijia Gao, Beibei Xu, Zijia Chen, Zhengjie Wang*

College of Electronic and Information Engineering, Shandong University of Science and Technology, Qingdao, China
Email: 3408929550@qq.com, 2076363523@qq.com, 19511699171@qq.com, fhj8ys@qq.com, 2945217990@qq.com, *cieewangzj@163.com

How to cite this paper: Li, H.T., Lu, Q.M., Gao, H.J., Xu, B.B., Chen, Z.J. and Wang, Z.J. (2026) A Review of Pedestrian Recognition Based on Millimeter-Wave Radar and Video Fusion. *Journal of Computer and Communications*, **14**, 103-118.
https://doi.org/10.4236/jcc.2026.144005

Received: April 5, 2026
Accepted: April 19, 2026
Published: April 22, 2026

Abstract

Person identification serves as a core supporting technology in public security, intelligent home applications, and person tracking in different environments. Achieving stable identity matching across environments and locations has become a critical research requirement. Single-modal identification technologies suffer from inherent limitations. The cross-modal fusion method based on millimeter-wave radar and video realizes the complementary advantages of the two modalities. It can effectively support person identification across scenarios, becoming a research hotspot in complex open scenarios. This paper systematically reviews the research progress and core technical systems in this field. Firstly, it sorts out the information extraction methods of millimeter-wave radar features and video visual features, summarizing the extraction ideas of key features such as range-Doppler, micro-Doppler, point cloud, and skeleton. Secondly, it summarizes three fusion strategies (data-level, feature-level, decision-level) and two core alignment methods (spatio-temporal alignment, feature distribution alignment). Then, it introduces typical applications, including intelligent security monitoring, human-computer interaction authentication, and multi-target tracking. It analyzes current challenges such as complex environmental interference, cross-location spatio-temporal alignment, feature distribution shift, and scarcity of large-scale data. Finally, it discusses future research directions, including scene-invariant feature learning, cross-domain alignment optimization, unified feature space construction, and few-shot generalization, providing a comprehensive reference for person identification research.

Keywords

Information Feature Extraction, Feature Fusion and Alignment, Cross-Modal Recognition and Matching

1. Introduction

With the rapid development of intelligent security, intelligent transportation, and public service technologies, person identification has become a core supporting technology in various fields [1]. Systems capable of stably completing person identification in different environments and locations are increasingly important. Traditional identification mostly relies on single-modal sensing in fixed scenarios and devices, which makes it difficult to meet the identity verification and target retrieval requirements across locations, devices, and environments in squares, stations, communities, buildings, etc. Person identification systems based on different sensing modalities have been continuously developed and applied. At present, mainstream identification systems can be mainly divided into: wearable sensor-based systems, computer vision-based systems, millimeter-wave radar-based systems, and millimeter-wave radar and video fusion-based systems.

Wearable sensor-based identification systems require users to wear devices integrated with sensors, such as fingerprints and heart rate, to collect biometric features [2]. However, wearable devices are inconvenient to use, high in maintenance costs, and cannot achieve non-cooperative cross-scenario identification, limiting their wide application in public areas. Unlike the cooperative identification that relies on active user cooperation in wearable sensor systems, the millimeter-wave radar and video fusion system features inherent non-cooperative sensing, with no need for on-body devices or active cooperation from targets. This characteristic is critical for identity verification in unconstrained public security scenarios such as squares, railway stations, and communities.

Computer vision-based identification systems complete identity verification by collecting visual features such as faces and gaits through cameras [3] [4]. RGB cameras can capture facial textures, and RGB-D sensors can obtain 3D structures [1], with high accuracy in ideal environments. However, visual methods have weak cross-scenario generalization ability, are easily affected by illumination, angle, occlusion, and clothing changes, and their performance decreases significantly in open environments such as squares and stations [5], making it difficult to support cross-location person identity matching and identification.

Millimeter-wave radar-based identification systems realize identity perception by analyzing radar echo signals caused by human movements, with advantages of non-cooperation, occlusion resistance, illumination independence, and strong privacy protection. Moreover, identity features such as gaits and micro-movements are more consistent across scenarios. However, a single radar modality has limited feature dimensions, making it difficult to achieve high-precision identity discrimination in long-distance and multi-target scenarios.

Millimeter-wave radar and video fusion-based person identification systems fully combine the complementary advantages of millimeter-wave radar and video. Millimeter-wave radar can penetrate obstacles, is not limited by illumination, and can capture 3D information with strong cross-scenario consistency, such as distance, speed, and gait trajectories, while protecting privacy [5] [6]; video can pro-

vide rich visual details such as faces, contours, and clothing [7]. Through dual-modal fusion, the system not only retains the high discrimination of visual features but also obtains the strong robustness of radar features, significantly improving the stability and accuracy of cross-environment, cross-location, and cross-device person identification. It is an ideal solution to realize cross-scenario retrieval and identity confirmation of personnel in open areas [8].

To date, remarkable progress has been made in the research on millimeter-wave radar and video fusion for person identification, but there is still a lack of comprehensive reviews specifically targeting this fusion technology. This paper systematically summarizes the latest research results on person identification methods based on millimeter-wave radar and video fusion. Firstly, it introduces the overall framework of the fusion identification system, including information feature extraction, feature fusion and alignment, and cross-modal recognition and matching algorithms. Then, it analyzes and summarizes typical application scenarios of the fusion system. Next, it discusses the current limitations and unsolved problems of fusion-based person identification and proposes potential future research directions. Finally, it concludes the full text to provide a reference for subsequent research in this field.

This paper is divided into five parts: The first part describes the development and application of person identification technologies based on different systems, emphasizing the advantages of millimeter-wave radar and video fusion systems. The second part details the overall framework and key technologies of the fusion identification system, including information feature extraction, feature fusion and alignment, and cross-modal recognition and matching algorithms. The third part introduces typical application scenarios of the fusion system and analyzes their performance characteristics. The fourth part points out the limitations and future research directions of fusion-based person identification. The fifth part presents the conclusions.

2. Architecture of Millimeter-Wave Radar and Video Fusion Person Identification System

The basic principle of person identification based on millimeter-wave radar and video fusion is as follows. Millimeter-wave radar and video sensors respectively, collect relevant information about target persons. The algorithms preprocess raw data to extract effective features with scene invariance, eliminate modal differences and scenario shifts through feature fusion and alignment, and finally complete identity verification through recognition and matching algorithms [1]. The general framework of the fusion identification system mainly includes three key stages, as shown in **Figure 1**.

- Information feature extraction: In this stage, raw data of target persons are collected through millimeter-wave radar and video sensors. The radar transmits frequency-modulated continuous waves and receives echo signals to extract 3D features such as distance, speed, and angle. Video sensors capture image

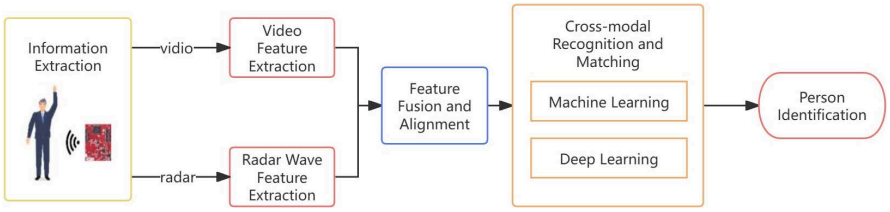

Figure 1. Architecture of person identification based on millimeter-wave radar and video fusion.

sequences to extract visual features such as facial textures, gait contours, and skeletal structures.

- Feature fusion and alignment: to remove noise and redundant information, the preprocessing of the extracted heterogeneous features is applied. This method gives the complementary advantages of the two modalities through fusion strategies such as data-level, feature-level, and decision-level. Meanwhile, alignment methods to reduce modal differences and map features to a unified feature space are implemented.
- Cross-modal recognition and matching: the system feeds the fused features into classification or matching algorithms to complete identity verification. A variety of machine learning and deep learning algorithms can be used, and the output result is the identity label of the target person.

These three stages constitute the general framework of person identification based on millimeter-wave radar and video fusion, providing support for the development of robust and reliable identification systems.

2.1. Feature Extraction

Feature extraction is the basis of fusion identification, aiming to fully capture identity-related features from radar and video data to lay the foundation for subsequent fusion and matching, as shown in Figure 2.

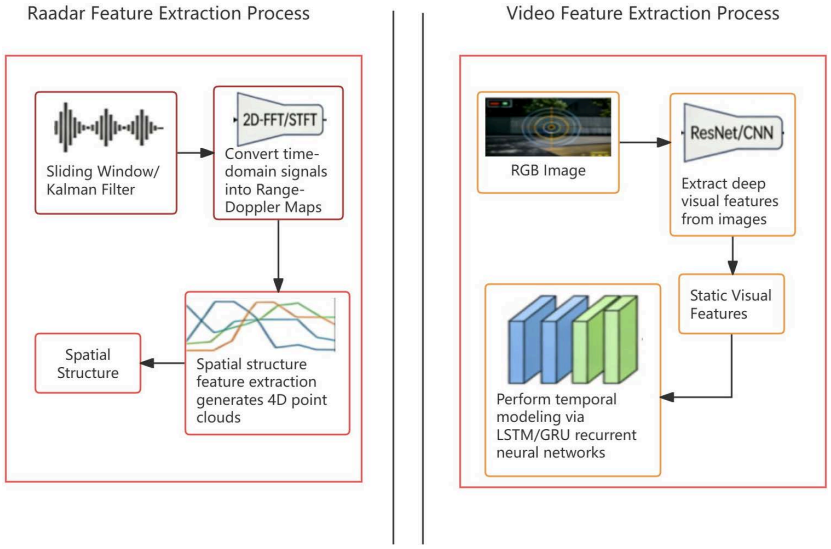

Figure 2. Feature extraction processes of the two modalities.

2.1.1. Radar Feature Extraction

Millimeter-wave radar extracts features based on echo signals, mainly including range-Doppler features, spatial structure features, and motion trajectory features [6].

The radar data collection usually requires the target to move naturally within the sensing range [2] to obtain sufficient echo information. Typical radar feature extraction applications involve various feature types, data forms, and sample sizes, as shown in **Table 1**.

Table 1. Typical applications of radar data acquisition and feature extraction.

Feature Type	Data Form	Sample Size	Radar Parameters	Reference
Contour features, motion trajectory	Voxelized point cloud	5 action categories, 3,200 samples	77 GHz FMCW radar	[9] [10]
MicroDoppler features, range-angle features	Timefrequency spectrogram	Multiple users, 1000+ samples	24 GHz SIMO radar	[11]
Vital sign features	Time series signal	20 volunteers	76 - 81 GHz FMCW radar	[2]
Radar point cloud features, 3D structural features	Coordinate matrix	58 volunteers	IWR6843ISKODS radar	[7]

Common radar feature extraction techniques include two-dimensional Fast Fourier Transform (2D-FFT), Multiple Signal Classification (MUSIC) algorithm, and point cloud voxelization. 2D-FFT converts time-domain signals into Range-Doppler Maps (RDM) to extract distance and speed features [9]. For example, Cao *et al.* [3] used 2D-FFT to process radar echo signals to obtain range-Doppler map sequences containing gait motion information; the MUSIC algorithm is used for high-precision angle estimation to generate Angle-Time Maps (ATM) to supplement spatial orientation features; point cloud voxelization converts sparse point clouds into regular voxel grids to facilitate feature extraction by deep learning models. Luo *et al.* [6] adopted a sliding window-based method to convert single-frame point clouds into multi-frame time-series data, retaining the temporal motion features of skeletal postures.

2.1.2. Video Feature Extraction

Video feature extraction focuses on identity-related visual information, such as facial features, gait features, and skeletal features [7]. Facial features include texture, contour, and key point information, extracted through face detection and feature encoding methods [12]; gait features originate from motion patterns such as walking posture and step frequency [10], usually obtained by analyzing image sequences or contour sequences [3] [4]; skeletal features extract 3D coordinates of human joints through pose estimation algorithms [6] [13].

Typical video feature extraction applications involve various feature types and processing methods, as shown in **Table 2**.

- Common data forms include Range-Doppler Spectrograms (RDM), 4D point clouds (x, y, z, v), time-frequency spectrograms, etc. Extraction technologies mainly include the following three categories:

Table 2. Typical applications of video data acquisition and feature extraction.

Feature Type	Processing Method	Data Source	Sample Size	Reference
RGB texture features, facial keypoint features	CNN-based feature extraction	RGB camera	58 volunteers × multiple images	[7]
Gait Energy Image (GEI), motion contour features	Spatial attention module	RGB camera	121 subjects × 8 views	[4]
Motion sequence features, visual semantic features	Transformer-based temporal modeling	RGB camera	Multiple sequences × 30+ frames	[8]
Facial identity features, gait motion features	Knowledge distillation encoding	RGB camera	36 subjects × 5000 + face-gait pairs	[12]

- Signal transformation technology converts time-domain signals into range-Doppler spectrograms through 2D-FFT to capture target motion features [14] and generates time-frequency spectrograms using STFT to extract micro-Doppler features [15] [16].
- Spatial feature extraction processes sparse point clouds using networks such as PointNet and DGCNN to capture 3D human structure and contour topology [7]. It also achieves high-precision angle estimation through the MUSIC algorithm to generate ATM [17].
- Temporal feature mining converts single-frame point clouds into time series using a sliding window mechanism to capture dynamic patterns of gaits and actions [18] and optimizes trajectory features through Kalman filtering to improve stability [19].

2.2. Feature Fusion and Alignment

Feature fusion and alignment are key elements for solving the heterogeneity of millimeter-wave radar and video data, directly affecting the performance of the fusion identification system [20]. The specific implementation is shown in **Figure 3**.

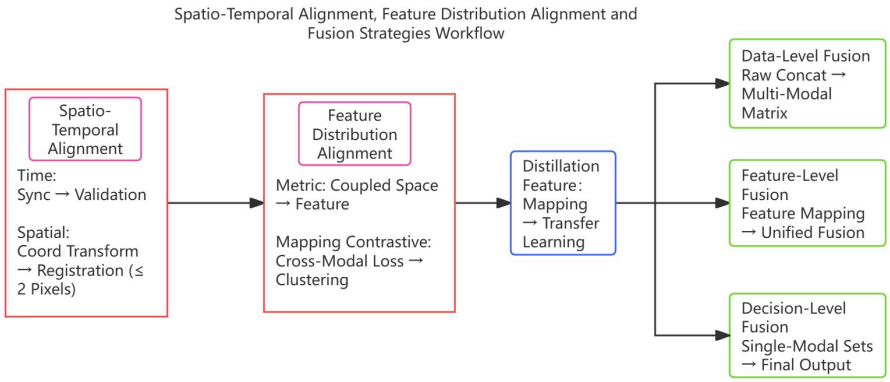

Figure 3. Implementation flow of spatio-temporal alignment, feature distribution alignment, and fusion strategies.

2.2.1. Fusion Strategies

According to different fusion stages, fusion strategies can be divided into data-

level fusion, feature-level fusion, and decision-level fusion [5].

Data-level fusion strategy directly integrates raw radar and video data before feature extraction. For example, Zong *et al.* [19] spliced radar range-angle information with video pixel data to form a multimodal data matrix, which is input into the fusion detection network. This strategy retains the most original information but has high requirements for data synchronization and noise reduction.

The feature-level fusion strategy integrates extracted radar and video features in a unified feature space, which is the most widely used fusion strategy in current research. For example, Liu *et al.* [7] designed a cross-modal similarity estimation method to fuse radar point cloud features and video 2D image features. Shi *et al.* [4] fused radar micro-motion features and video gait appearance features in a multi-scale feature space. Chen *et al.* [8] adopted a dual contrast learning framework to fuse radar Doppler features and video motion features. Feature-level fusion balances information retention and computational efficiency, and can adaptively adjust feature weights according to scenario changes.

Decision-level fusion strategy combines the recognition results of radar and video single-modal systems to obtain the final identity judgment [15]. It has strong robustness to single-modal failure but relies on the accuracy of single-modal decision results.

The three fusion strategies present significant trade-offs between computational cost and system latency. Data-level fusion entails processing raw heterogeneous data, yielding high computational complexity and maximum system latency. Feature-level fusion only integrates the extracted feature vectors, with moderate computational cost and latency, balancing information integrity and real-time performance. Decision-level fusion merely conducts posterior fusion on the recognition outcomes of single-modal systems and delivers minimal computational overhead and the lowest system latency, thus being better suited for application scenarios with stringent real-time requirements.

2.2.2. Alignment Methods

Alignment methods aim to eliminate spatio-temporal asynchrony and feature distribution differences between radar and video modalities [21], including spatio-temporal alignment and feature distribution alignment.

Spatio-temporal alignment ensures that the features of the two modalities correspond to the same target at the same time [6]. Time alignment is usually realized through sensor timestamp synchronization [7]. Spatial alignment uses calibration methods to map radar spatial coordinates (distance, angle) to video image coordinates to achieve spatial registration of targets. For example, Yang *et al.* proposed the BiCro method [22], which corrects noisy correspondences through bidirectional cross-modal similarity consistency to improve alignment accuracy.

Feature distribution alignment reduces the distribution gap between radar and video features. Metric learning is a common method. Wang *et al.* [20] learned a coupled feature space to map heterogeneous features to a unified space. Wei *et al.* [21] proposed a universal weighted metric learning method to adaptively adjust

feature weights. Liong *et al.* [23] adopted deep coupled metric learning to realize nonlinear mapping of features. In addition, contrastive learning and knowledge distillation are also widely used in alignment tasks [12]. Chen *et al.* [8] transferred video facial identity knowledge to radar gait features through knowledge distillation to achieve semantic alignment.

2.3. Cross-Modal Recognition and Matching Algorithms

Cross-modal matching algorithms [24] and recognition [25] complete identity verification based on fused and aligned features, which can be divided into machine learning algorithms and deep learning algorithms. Traditional machine learning relies on manually designed features, with lightweight models, fast training speed, good compatibility with small sample data, and strong interpretability. However, it has limited feature expression ability in complex scenarios and vulnerable generalization performance, and is suitable for classification and regression tasks with small data volume and clear features. Deep learning can automatically learn deep abstract features from massive data, with higher accuracy in complex data processing such as images and radar signals, but has large model parameters, a long training time, requires a large amount of labeled data, poor interpretability, and is prone to overfitting. Overall, machine learning is more advantageous in small-sample and low-dimensional scenarios; deep learning has significantly better feature extraction and recognition performance in high-dimensional and nonlinear data tasks such as millimeter-wave radar and video fusion.

2.3.1. Machine Learning Algorithms

Machine learning algorithms mainly use traditional classification and matching methods to process fused features [20]. Common algorithms include Dynamic Time Warping (DTW), K-Nearest Neighbor (KNN), Hidden Markov Model (HMM), and Random Forests.

DTW is used for matching time-series features such as gaits and motion trajectories, calculating the similarity between test sequences and reference templates through dynamic programming, suitable for radar motion feature matching. KNN performs identity classification based on feature similarity [9]. For example, Guo *et al.* [2] used KNN to classify fused vital sign features with an accuracy of over 85%. HMM models the temporal dependence of features, suitable for dynamic gesture and gait-based recognition [10]. Random Forest performs classification by integrating multiple decision trees and can handle high-dimensional fused features [11]. However, machine learning algorithms have a limited ability to process complex heterogeneous features and perform worse than deep learning algorithms on large-scale datasets.

2.3.2. Deep Learning Algorithms

Deep learning algorithms have powerful feature learning and pattern recognition capabilities, making them the mainstream algorithms in current fusion identification [7] [12]. Common algorithms include Convolutional Neural Networks (CNN),

Long Short-Term Memory (LSTM), and Transformer.

CNN is used to extract spatial features from voxelized point clouds and multi-modal feature maps [14]. For example, Shi *et al.* [4] used a deep CNN to fuse radar time-Doppler spectrograms and video gait energy images. Luo *et al.* [6] used CNN to extract spatial features of radar point clouds. LSTM is suitable for processing time-series features such as gait sequences and radar signal sequences. Shan *et al.* [12] used LSTM to model the temporal dependence of gait features to improve recognition robustness. Transformer captures global feature dependencies through a self-attention mechanism, suitable for cross-modal feature matching [18]. Chen *et al.* [8] used Transformer-based networks for temporal modeling and instance discrimination to enhance the discriminative ability of fused features.

Some innovative algorithms were proposed for these applications. Yao *et al.* [25] proposed the semi-supervised recognition system mmSignature, achieving 96.3% accuracy with a small amount of labeled data. Shao *et al.* [18] realized fast action recognition based on point cloud sequence learning with inference time less than 0.2 ms. Deep learning algorithms can also be combined with multi-task learning, transfer learning, and other technologies to improve performance. For example, Shan *et al.* [12] used knowledge distillation for transfer learning, and Deng *et al.* [26] used synthetic data to enhance model generalization ability.

In summary, deep learning delivers superior performance in high-dimensional, heterogeneous cross-modal feature learning and complex-scenario recognition. By contrast, traditional machine learning models still present more prominent practical advantages for resource-constrained edge computing devices, few-shot annotation scenarios, and low-latency lightweight deployment, owing to their lightweight architecture, low computational cost, and fast inference speed.

3. Typical Applications

This section introduces several typical application systems of person identification based on millimeter-wave radar and video fusion, and analyzes their performance characteristics from multiple perspectives. These studies are divided into the following categories according to application scenarios: intelligent security monitoring, human-computer interaction identity authentication, and multi-target tracking identification.

3.1. Intelligent Security Monitoring

This application aims to achieve real-time person identification and pedestrian re-identification in public places such as shopping malls, stations, and communities [27], ensure public safety, and provide cross-environment identity confirmation and trajectory tracking of personnel [6]. The core requirements are high precision and strong robustness to complex environments.

Cao *et al.* [3] proposed a cross vision-RF gait re-identification system based on low-cost RGB-D cameras and millimeter-wave radar, fusing gait motion features extracted by radar and visual gait features obtained by cameras, achieving about

92.5% top-1 accuracy and 97.5% top-5 accuracy among 56 volunteers, and stably identifying targets even in multi-person scenarios. Liu *et al.* [7] designed the Mission system, which detects targets through radar in camera-restricted areas and identifies person images from the camera network, achieving 85% top-1 accuracy and 90% top-5 accuracy among 58 volunteers, suitable for security monitoring scenarios.

The core challenge of intelligent security monitoring is to cope with environmental interference such as crowd density, occlusion, and illumination changes. Future research needs to focus on improving the real-time performance and multi-target discrimination ability of the system.

3.2. Human-Computer Interaction Identity Authentication

This application combines person identification with human-computer interaction, realizing identity authentication while completing gesture and action control, and is applied to smart homes, smart vehicles, and other scenarios. The core requirements are fast response and non-contact operation. Janakaraj *et al.* [28] proposed the STAR system, realizing simultaneous tracking and recognition of persons using millimeter-wave radar and deep learning, fusing radar tracking information and video identity features, achieving real-time recognition in human-computer interaction scenarios with high stability. Huang *et al.* [11] proposed the RPCRS system that extracted human activity characteristics from millimeter-wave radar signals and combined them with video identity verification to achieve safe human-computer interaction, with recognition accuracy over 90%. The MatchAnything method, proposed by He *et al.* [29], is capable of matching diverse image data, making it applicable for cross-modal behavior recognition, where it can also be utilized to process radar signals to enhance recognition precision. Yao *et al.*'s mmSignature system [30] realizes identity verification for interactive scenarios such as device unlocking and personalized services through semi-supervised learning, with an accuracy of 96.3%.

The core challenge of human-computer interaction identity authentication is to balance real-time performance and accuracy. Future research needs to design lightweight fusion models to reduce computational delays.

3.3. Multi-Target Tracking Identification

This application aims to continuously maintain identity and stably identify multiple moving personnel in complex and dense scenarios such as stations, squares, and parks. The core requirement is to distinguish different personnel identities and maintain recognition consistency even under multi-person interference, occlusion, and staggered movement. Luo *et al.* [6] adopted the CNN-BiGRU architecture to fuse radar point clouds and video joint information, realizing multi-target skeletal positioning and identity discrimination, with a multi-target identity identification accuracy of 89%. Cao *et al.* [3] fused radar gait features and video visual features, and can stably match target identities even in multi-person parallel movement scenarios, effectively avoiding target confusion and identity jumps.

Shao *et al.* [18] realized real-time multi-target identity tracking and discrimination based on rapid modeling of radar point cloud sequences, maintaining high frame rate and identity consistency in complex crowd scenarios.

The core challenge of multi-target tracking identity identification is feature confusion and identity association errors caused by occlusion interference and target staggering. Future optimization of the multi-target association mechanism and cross-modal feature matching strategy is needed to improve the stability of identity identification in dense scenarios.

3.4. Summary

Table 3 briefly summarizes several mainstream applications, listing the goals and key elements of four typical application categories.

Table 3. Key features of four typical application categories.

Application Type	Goal	Highlights	Reference
Intelligent security monitoring	Real-time identification in public places	High precision, anti-interference	[1] [3] [7]
Human-computer interaction authentication	Authentication during interaction	Fast response, non-contact	[11] [28]
Multi-target tracking identification	Simultaneous identification of multiple targets	Occlusion resistance, consistent tracking	[6] [18]

4. Current Challenges

4.1. Complex Environmental Interference

Most current research experiments are conducted in laboratory environments with less interference [2] [7]. However, in practical applications, the system is often affected by factors such as crowd density, occlusion, illumination changes, and electromagnetic interference [1]. For example, in dense crowd scenarios, radar signals are easily mixed with echoes from multiple targets, and video images are severely occluded, leading to a significant decrease in recognition accuracy [6]. In addition, severe weather such as rain and fog will attenuate radar signals and blur video images, further affecting system performance [5].

4.2. Limited Modal Alignment Accuracy

Although existing alignment methods have made progress, there are still gaps in spatio-temporal synchronization and feature distribution consistency [20] [23]. Spatio-temporal asynchrony between radar and video sensors may lead to feature misalignment of the same target [6]. Feature distribution differences caused by modal heterogeneity [29] make it difficult to fully integrate complementary information, limiting the improvement of recognition accuracy [12]. Especially in dynamic scenarios, target movement will exacerbate alignment difficulties [8].

4.3. Scarcity of Large-Scale Data

The collection and annotation of millimeter-wave radar and video fusion data are

time-consuming and labor-intensive. Cross-scenario data collection and annotation are costly and privacy-sensitive, making it difficult to construct large-scale standardized datasets [26]. Currently, publicly available millimeter-wave radar and video fusion datasets for person identification are extremely scarce. Two representative datasets are as follows. The mmBody dataset provides synchronized millimeter-wave radar point clouds and RGBD video data, covering 20 subjects, 100 motions, and 7 scenes, serving as a commonly used multimodal benchmark for human perception [31]. Such public datasets generally suffer from limitations, including a small number of subjects, single-scene settings, and insufficient sample sizes, which still cannot meet the requirements of large-scale cross-scenario model training. Radar data require professional equipment and technicians for collection, and video data involve privacy issues, making it difficult to build large-scale annotated datasets [16]. Although some studies use synthetic data to supplement training samples [26], there is a distribution gap between synthetic data and real data, affecting model generalization ability [30].

4.4. Privacy and Ethical Risks

Although millimeter-wave radar data is privacy-preserving and avoids direct collection of sensitive human visual information, fusion of radar data and video data forms a multimodal data linkage that can precisely associate individual identities, thereby introducing new privacy leakage and ethical risks. Cross-modal data association and matching may break through the privacy protection boundaries of a single modality. Without sufficient authorization, this can easily lead to the abuse of identity information and continuous tracking, violating the principles of data minimization and privacy compliance. This represents a critical ethical and compliance challenge that must be addressed in the practical deployment of fusion technologies.

5. Future Research Directions

5.1. Enhance Environmental Adaptability and Learn Scene-Invariant Features

To improve anti-interference ability, it is important to optimize feature extraction algorithms [3]. For example, we may use attention mechanisms to focus on extracting "illumination-independent, angle-independent, occlusion-robust" identity features, and use attention mechanisms, domain adaptation, and data enhancement to improve cross-environment generalization ability and suppress environmental noise [4] [8]. We also design a multi-view sensor layout to reduce occlusion impact [32]. In addition, we may enhance model robustness through domain adaptation and data enhancement technologies [26].

5.2. Improve Modal Alignment Accuracy

We may develop more accurate spatio-temporal calibration methods to achieve millisecond-level time synchronization and pixel-level spatial registration [6]. At the same time, we can explore advanced feature alignment technologies such as

cross-modal contrastive learning and generative adversarial networks to reduce feature distribution differences [12]. For dynamic scenarios, it is important to design adaptive alignment strategies that can be adjusted in real time according to target movement [8].

5.3. Expand Data Resources and Cross-Scenario Generalization

High-quality synthetic data may be generated to supplement real data [26] to reduce dependence on large-scale annotated data and achieve rapid adaptation to new locations and environments with a small number of samples. For example, [33] optimizes the video-to-radar conversion model to improve the authenticity of synthetic radar data [16]. We may use federated learning to achieve data sharing under privacy protection, expand the training data scale, and explore semi-supervised and unsupervised learning algorithms to reduce dependence on annotated data [29].

5.4. Unified Feature Space Construction

Unified feature space construction is a crucial research direction, as it can map data from multiple modalities into a common feature space. As a result, it reduces the discrepancies among different modalities, enables more diverse applications, and allows the use of more network models. Unified feature space construction can rely on modern architectures, including cross-modal Transformers, dual-branch contrastive learning frameworks, and deep cross-modal metric learning networks. It realizes global cross-modal feature interaction via self-attention mechanisms and narrows the feature distance of homologous targets through contrastive learning, ultimately achieving precise alignment and efficient representation of heterogeneous radar and video features in a unified latent space. This serves as the core technical pathway to address modal heterogeneity.

6. Conclusions

This paper comprehensively reviews the research status of person identification methods based on millimeter-wave radar and video fusion. Firstly, it reviews common person identification systems, dividing them into wearable sensor-based, computer vision-based, millimeter-wave radar-based, and millimeter-wave radar and video fusion-based systems, highlighting the advantages of dual-modal fusion. Then, it introduces the overall framework of the fusion identification system, including information feature extraction, feature fusion and alignment, and cross-modal recognition and matching algorithms, and elaborates on key technologies and implementation methods in detail. Next, it reviews the latest research on fusion-based person identification, dividing it into three typical application scenarios: intelligent security monitoring, human-computer interaction identity authentication, and multi-target tracking identification, and analyzes the implementation process and performance characteristics of each system. Finally, it discusses the limitations and existing problems of current research and proposes future re-

search directions combined with research trends.

Person identification based on millimeter-wave radar and video fusion combines the advantages of the two modalities, overcomes the limitations of single-modal identification, and has broad application prospects in intelligent security, human-computer interaction, and other fields. With the continuous advancement of sensor technology and artificial intelligence algorithms, future fusion identification systems will achieve higher accuracy, greater robustness, and better real-time performance, providing more reliable technical support for various intelligent applications. This paper can provide a reference for researchers' follow-up research and help develop more efficient and practical person identification systems.

Funding

The work is funded by the Foundation of the Innovation and Entrepreneurship Training Program for College Students (X202510424033).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Yang, Q., Quan, Z., Li, J., Jiang, T., Deng, Z., Qiu, X., *et al.* (2025) A Survey of Pedestrian Re-Identification Based on Millimeter Wave Radar and Vision Fusion. *Journal of Computer and Communications*, **13**, 64-80. https://doi.org/10.4236/jcc.2025.136005

[2] Guo, J., Wei, J., Xiang, Y. and Han, C. (2024) Millimeter-Wave Radar-Based Identity Recognition Algorithm Built on Multimodal Fusion. *Sensors*, **24**, Article 4051. https://doi.org/10.3390/s24134051

[3] Cao, D., Liu, R., Li, H., Wang, S., Jiang, W. and Lu, C.X. (2022) Cross Vision-RF Gait Re-Identification with Low-Cost RGB-D Cameras and Mmwave Radars. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, **6**, 1-25. https://doi.org/10.1145/3550325

[4] Shi, Y., Du, L., Chen, X., Liao, X., Yu, Z., Li, Z., *et al.* (2023) Robust Gait Recognition Based on Deep CNNs with Camera and Radar Sensor Fusion. *IEEE Internet of Things Journal*, **10**, 10817-10832. https://doi.org/10.1109/jiot.2023.3242417

[5] Wang, S., Mei, L., Liu, R., Jiang, W., Yin, Z., Deng, X., *et al.* (2025) Multi-Modal Fusion Sensing: A Comprehensive Review of Millimeter-Wave Radar and Its Integration with Other Modalities. *IEEE Communications Surveys & Tutorials*, **27**, 322-352. https://doi.org/10.1109/comst.2024.3398004

[6] Luo, Y., He, Y., Li, Y., Liu, H., Wang, J. and Gao, F. (2025) A Sliding Window-Based CNN-BiGRU Approach for Human Skeletal Pose Estimation Using Mmwave Radar. *Sensors*, **25**, Article 1070. https://doi.org/10.3390/s25041070

[7] Liu, R., Yao, T., Shi, R., Mei, L., Wang, S., Yin, Z., *et al.* (2024) Mission: mmWave Radar Person Identification with RGB Cameras. *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*, Hangzhou, 4-7 November 2024, 309-321. https://doi.org/10.1145/3666025.3699340

[8] Chen, Y. and Cheng, K. (2024) BiCLR: Radar-Camera-Based Cross-Modal Bi-Contrastive Learning for Human Motion Recognition. *IEEE Sensors Journal*, **24**, 4102-

4119. https://doi.org/10.1109/jsen.2023.3344789

[9] Singh, A.D., Sandha, S.S., Garcia, L. and Srivastava, M. (2019) RadHAR: Human Activity Recognition from Point Clouds Generated through a Millimeter-Wave Radar. *Proceedings of the* 3*rd ACM Workshop on Millimeter-Wave Networks and Sensing Systems*, Los Cabos, 25 October 2019, 51-56. https://doi.org/10.1145/3349624.3356768

[10] Li, Z., Le Kernec, J., Abbasi, Q., Fioranelli, F., Yang, S. and Romain, O. (2023) Radar-based Human Activity Recognition with Adaptive Thresholding towards Resource Constrained Platforms. *Scientific Reports*, **13**, Article No. 3473. https://doi.org/10.1038/s41598-023-30631-x

[11] Huang, T., Liu, G., Li, S. and Liu, J. (2023) RPCRS: Human Activity Recognition Using Millimeter Wave Radar. 2022 *IEEE* 28*th International Conference on Parallel and Distributed Systems* (*ICPADS*), Nanjing, 10-12 January 2023, 122-129. https://doi.org/10.1109/icpads56603.2022.00024

[12] Shan, L., Zhang, R., Chilukoti, S.V., Zhang, X., Lee, I. and Hei, X. (2024) IdentityKD: Identity-Wise Cross-Modal Knowledge Distillation for Person Recognition via Mmwave Radar Sensors. *Proceedings of the* 6*th ACM International Conference on Multimedia in Asia*, Auckland, 3-6 December 2024, 1-7. https://doi.org/10.1145/3696409.3700254

[13] Yang, Z. and Yang, Y. (2025) Wisk: A Lightweight Multimodal Human Activity Recognition Method Based on WiFi Channel State Information and Video Skeleton Images. *Concurrency and Computation*: *Practice and Experience*, **37**, e70383. https://doi.org/10.1002/cpe.70383

[14] Cao, P., Xia, W., Ye, M., Zhang, J. and Zhou, J. (2018) Radar-ID: Human Identification Based on Radar Micro-Doppler Signatures Using Deep Convolutional Neural Networks. *IET Radar*, *Sonar & Navigation*, **12**, 729-734. https://doi.org/10.1049/iet-rsn.2017.0511

[15] Hafner, F.M., Bhuiyan, A., Kooij, J.F.P. and Granger, E. (2019) RGB-Depth Cross-Modal Person Re-Identification. 2019 16*th IEEE International Conference on Advanced Video and Signal Based Surveillance* (*AVSS*), 18-21 September 2019, 1-8. https://doi.org/10.1109/avss.2019.8909838

[16] Ahuja, K., Jiang, Y., Goel, M. and Harrison, C. (2021) Vid2Doppler: Synthesizing Doppler Radar Data from Videos for Training Privacy-Preserving Activity Recognition. *Proceedings of the* 2021 *CHI Conference on Human Factors in Computing Systems*, Yokohama, 8-13 May 2021, 1-10. https://doi.org/10.1145/3411764.3445138

[17] Xie, S., Wang, C., Yang, X., Wan, Y., Zeng, T. and Liu, Z. (2022) Millimeter-Wave Radar Target Detection Based on Inter-Frame DBSCAN Clustering. 2022 *IEEE* 22*nd International Conference on Communication Technology* (*ICCT*), Nanjing, 11-14 November 2022, 1703-1708. https://doi.org/10.1109/icct56141.2022.10072664

[18] Shao, T., Du, Z., Li, C., Wu, T. and Wang, M. (2024) Fast Human Action Recognition via Millimeter Wave Radar Point Cloud Sequences Learning. *Proceedings of the* 33*rd ACM International Conference on Information and Knowledge Management*, Boise, 21-25 October 2024, 2024-2033. https://doi.org/10.1145/3627673.3679787

[19] Zong, M., Wu, J., Zhu, Z. and Ni, J. (2024) A Method for Target Detection Based on Mmw Radar and Vision Fusion. arXiv:2403.16476.

[20] Wang, K., He, R., Wang, W., Wang, L. and Tan, T. (2013) Learning Coupled Feature Spaces for Cross-Modal Matching. 2013 *IEEE International Conference on Computer Vision*, Sydney, 1-8 December 2013, 2088-2095. https://doi.org/10.1109/iccv.2013.261

[21] Wei, J., Xu, X., Yang, Y., Ji, Y., Wang, Z. and Shen, H.T. (2020) Universal Weighting

Metric Learning for Cross-Modal Matching. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 6420-6429. https://doi.org/10.1109/cvpr42600.2020.01302

[22] Yang, S., Xu, Z., Wang, K., You, Y., Yao, H., Liu, T., *et al.* (2023) BiCro: Noisy Correspondence Rectification for Multi-Modality Data via Bi-Directional Cross-Modal Similarity Consistency. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 19883-19892. https://doi.org/10.1109/cvpr52729.2023.01904

[23] Liong, V.E., Lu, J., Tan, Y. and Zhou, J. (2017) Deep Coupled Metric Learning for Cross-Modal Matching. *IEEE Transactions on Multimedia*, **19**, 1234-1244. https://doi.org/10.1109/tmm.2016.2646180

[24] Wei, J., Xu, X., Wang, Z. and Wang, G. (2021) Meta Self-Paced Learning for Cross-Modal Matching. *Proceedings of the 29th ACM International Conference on Multimedia*, Virtual Event, 20-24 October 2021, 1613-1621. https://doi.org/10.1145/3474085.3475451

[25] Yao, Y., Zhang, H., Xia, P., Liu, C., Geng, F., Bai, Z., *et al.* (2023) Mmsignature: Semi-Supervised Human Identification System Based on Millimeter Wave Radar. *Engineering Applications of Artificial Intelligence*, **126**, Article 106939. https://doi.org/10.1016/j.engappai.2023.106939

[26] Deng, K., Zhao, D., Han, Q., Zhang, Z., Wang, S., Zhou, A., *et al.* (2023) Midas: Generating mmWave Radar Data from Videos for Training Pervasive and Privacy-Preserving Human Sensing Tasks. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, **7**, 1-26. https://doi.org/10.1145/3580872

[27] Hafner, F.M., Bhuyian, A., Kooij, J.F.P. and Granger, E. (2022) Cross-Modal Distillation for RGB-Depth Person Re-Identification. *Computer Vision and Image Understanding*, **216**, Article 103352. https://doi.org/10.1016/j.cviu.2021.103352

[28] Janakaraj, P., Jakkala, K., Bhuyan, A., Sun, Z., Wang, P. and Lee, M. (2019) STAR: Simultaneous Tracking and Recognition through Millimeter Waves and Deep Learning. 2019 12*th IFIP Wireless and Mobile Networking Conference* (*WMNC*), Paris, 11-13 September 2019, 211-218. https://doi.org/10.23919/wmnc.2019.8881354

[29] He, X.Y., Yu, H., Peng, S.D., Tan, D.L., Shen, Z.H., Bao, H.J. and Zhou, X.W. (2025) MatchAnything: Universal Cross-Modality Image Matching with Large-Scale Pre-Training. arXiv:2501.07556v1.

[30] Vishwakarma, S., Li, W., Tang, C., Woodbridge, K., Adve, R. and Chetty, K. (2023) SimHumalator: An Open Source End-to-End Radar Simulator for Human Activity Recognition. *IEEE Aerospace and Electronic Systems Magazine*, **37**, 6-22. https://doi.org/10.1109/MAES.2021.3138948

[31] Chen, A., Wang, X., Shi, K., Huo, Y., Chen, J. and Ye, Q. (2025) Toward Weather-Robust 3D Human Body Reconstruction: Millimeter-Wave Radar-Based Dataset, Benchmark, and Multi-Modal Fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, **35**, 273-286. https://doi.org/10.1109/tcsvt.2024.3461960

[32] Choi, J., Hor, S., Yang, S. and Arbabian, A. (2025) Mvdoppler-Pose: Multi-Modal Multi-View Mmwave Sensing for Long-Distance Self-Occluded Human Walking Pose Estimation. 2025 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 10-17 June 2025, 1-8. https://doi.org/10.1109/cvpr52734.2025.02584

[33] Huan, S., Wang, Z., Wang, X., Wu, L., Yang, X., Huang, H., *et al.* (2023) A Lightweight Hybrid Vision Transformer Network for Radar-Based Human Activity Recognition. *Scientific Reports*, **13**, Article No. 17996. https://doi.org/10.1038/s41598-023-45149-5