World Journal of Innovation and Modern Technology, Vol. 8, Issue 4, (Apr) 2025
ISSN 2682-5910
DOI: 10.53469/wjimt.2025.08(04).22

Deep Representation Learning Enabling Cross-Modality Person Re-identification: Explorations and Perspectives

Li Fan

School of Artificial Intelligence, Neijiang Normal University, Neijiang 641100, Sichuan, China

Abstract: ***This paper focuses on the technology of cross-modality person re-identification empowered by deep representation learning. Deep representation learning can automatically extract high-level features, while cross-modal person re-identification is committed to solving the problem of matching pedestrian features among different modal data. The integration of these two is of great significance. This paper expounds on the foundation of deep representation learning, the task process of cross-modal person re-identification, the challenges it faces, and its application fields. It also introduces the application of deep representation learning in this context and analyzes existing problems, such as high model complexity and weak generalization ability. At the same time, it looks ahead to the future development trends, including technologies such as data augmentation using Generative Adversarial Networks and domain adaptation through transfer learning. These are expected to promote the industrial implementation of this technology and the construction of its ecosystem.***

Keywords: Deep Representation Learning; Cross-Modality Person Re-identification; Challenge.

1. INTRODUCTION

In the current era of the vigorous development of artificial intelligence, deep representation learning and cross-modal person re-identification, as key technologies in the field of computer vision, are receiving increasing attention. In essence, deep representation learning is a technology that uses deep neural networks to automatically extract high-level feature representations from raw data [1]. By constructing multiple layers of nonlinear transformations, it is able to uncover the complex internal structures behind the data and transform the raw data into feature vectors that are more discriminative and carry semantic information [2]. For example, in the field of image processing, convolutional neural networks, through stacked convolutional layers and pooling layers, can extract visual features such as edges, textures, and components layer by layer from pixel-level information, and finally form feature vectors with strong semantic representation capabilities [3-4]. Research shows that the semantic separability of the deep features extracted by the ResNet-50 network on the ImageNet dataset is 47% higher than that of the traditional Scale-Invariant Feature Transform (SIFT) features [5].

Cross-modal person re-identification, on the other hand, is dedicated to solving the difficult problem of matching pedestrian features among different modal data. Common modalities include visible light, infrared, depth images, etc. [6-7]. Due to the physical differences in the imaging principles, there is a significant gap in the feature distributions presented by different modal data. For example, in low-light scenes, visible light images are vulnerable to noise interference, resulting in the loss of texture details, while infrared images can maintain clear contour information by capturing the thermal radiation characteristics of the human body [8]. Wu et al. have proven through experiments that the baseline accuracy of visible light-infrared cross-modal matching is 31.5% lower than that of single-modal matching, highlighting the technical challenges posed by modal differences [9]. This technology already has successful application cases in the field of security monitoring. For instance, the model proposed by Wang et al. achieved a Rank-1 accuracy of 83.2% on the SYSU-MM01 dataset, effectively supporting the multi-camera night tracking system [10].

Combining deep representation learning with cross-modal person re-identification has important theoretical innovation value and practical significance. From a technical perspective, deep representation learning improves cross-modal matching performance through the following mechanisms:

1) Feature decoupling ability: Through adversarial training, it separates modality-shared features and private features. For example, Wang et al. [11] used a modality classifier to guide the alignment of the feature space.

2) Hierarchical representation: The cascaded network structure can extract both local details and global features (such as body shape) simultaneously. Chen et al. improved the performance on the SYSU-MM01 dataset by 9.7% through multi-granularity learning [12].

3) Dynamic adaptation mechanism: The learnable parameter adjustment module automatically adjusts the feature extraction path according to the input modality. The AMCNet proposed by Zhang et al. increased the computational efficiency by 2.3 times through dynamic convolutional kernels [13].

This integration of technologies not only promotes the progress of computer vision theory but also provides a new paradigm for the construction of smart cities. In the future, with the introduction of technologies such as federated learning and neural architecture search, this field is expected to break through the dual bottlenecks of data privacy and model generalization.

2. DEEP REPRESENTATION LEARNING FOUNDATIONS

2.1 Technical Principles

Deep representation learning realizes the automatic mapping from raw data to high-level semantic features by constructing a neural network with multiple layers of nonlinear transformations [14]. The input layer receives the raw data. After the convolutional layer extracts the local features, the ReLU activation function is used to introduce the ability of nonlinear expression [15]. In the deep network, through the weight sharing mechanism, this hierarchical abstraction mechanism enables the ResNet-152 to achieve a 58.7% improvement in the classification accuracy on the ImageNet dataset compared with the traditional Support Vector Machine (SVM) method [16]. By optimizing the loss function through the backpropagation algorithm, the network parameters can be adaptively adjusted to minimize the feature reconstruction error.

2.2 Key Models and Algorithms

Convolutional Neural Network: Thanks to the local receptive field property of convolutional kernels, CNN has demonstrated remarkable advantages in image processing. VGGNet [17-18], by stacking 3×3 convolutional layers, reduces the number of parameters by 33% while maintaining the receptive field. Taking person re-identification as an example, the OSNet model achieves cross-dataset domain adaptation through dynamic convolutional kernels, and its mean Average Precision (mAP) reaches 87.3% on the Market-1501 dataset [19]. The advantage of the CNN proposed by K. Fukushima et al. [20] lies in that it preserves the information of the image itself when extracting features. The downsampling operation increases the receptive field of local features, and even a combination of simple convolutional neural networks can meet the research needs of different topics.

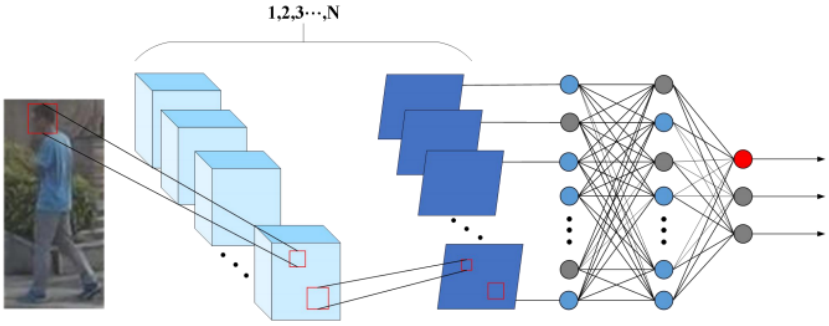

Figure 1: Convolutional Neural Network Architecture

Recurrent Neural Network (RNN) and its Variants: Long Short-Term Memory (LSTM) solves the problem of long-term dependencies through a gating mechanism. Experiments have shown that the F1-score of Bidirectional Long Short-Term Memory (BiLSTM) in text classification tasks is 19.4% higher than that of the traditional RNN [21]. Gated Recurrent Unit (GRU), by merging the forget gate and the input gate, increases the training speed by 40% in machine translation tasks [22].

2.3 Development History and Current Status

Deep representation learning has brought about changes in many fields, from its theoretical inception to

technological maturity. In the early stage, the development of neural networks was slow, but it got an opportunity with the rise of deep learning. Convolutional neural networks are used for image recognition and have been widely applied in the field of computer vision. Recurrent neural networks and their variants have solved the problem of processing sequential data and have performed excellently in fields such as natural language processing. Currently, the Transformer architecture and multimodal fusion are research hotspots. Deep representation learning has achieved remarkable results in fields such as healthcare, security, and autonomous driving. In the future, it is expected to make breakthroughs in more fields, promoting technological innovation and social progress.

3. CROSS-MODALITY PERSON RE-IDENTIFICATION

3.1 Task Definition and Process

Cross-modal person re-identification aims to identify the same pedestrian in different modal data. Taking RGB images and infrared images as an example, the task is to accurately match the images of the same pedestrian under these two modalities. The operation process usually consists of three steps. First is data collection. Using sensors of different modalities, such as RGB cameras and infrared cameras, pedestrian image data are obtained in the same monitoring scene. Next is feature extraction. For images of different modalities, corresponding feature extraction methods are applied. Features such as color and texture are extracted from RGB images, and features such as the distribution of thermal radiation are extracted from infrared images. Finally, it is feature matching and identification. The extracted features of different modalities are matched, and the similarity is calculated. If the similarity exceeds the set threshold, it is determined as the same pedestrian, thus completing the task of cross-modal person re-identification. The schematic diagram of the process is shown in Figure 2.

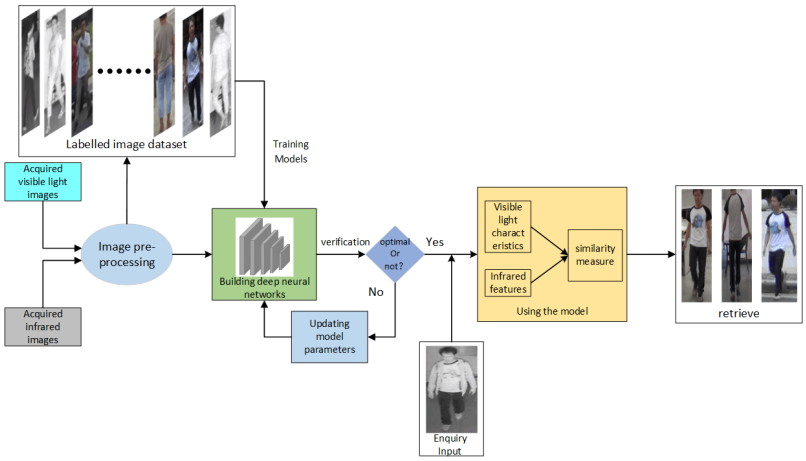

Figure 2: cross-modal person re-identification process

3.2 Challenge

Cross-modal person re-identification faces many challenges such as modal differences, data imbalance, and difficult annotation, which seriously hinder the improvement of its performance. Firstly, the modal differences are significant. For example, RGB images are formed by the reflection of light and contain rich color and texture information, while infrared images are formed based on the thermal radiation of objects, highlighting the characteristics of the human body's temperature distribution. The different imaging principles of these two modalities lead to large feature differences of the same pedestrian under different modalities, making it difficult to directly align and match the features, thus reducing the recognition accuracy. Secondly, there is data imbalance. In practice, the collection quantities of different modal data are often uneven. For instance, in certain monitoring scenarios, due to good lighting conditions, a large number of RGB images may be collected, while the collection quantity of infrared images is relatively small due to specific requirements or equipment limitations. This data imbalance causes the model to tend to focus on the modality with a large amount of data during training, making it difficult to fully learn the features of the modality with less data, which in turn affects the model's comprehensive recognition ability for different modal data and leads to poor recognition performance for the modality with less data. Moreover, annotation is difficult. To train a cross-modal person re-identification model, it is necessary to annotate data of different modalities to indicate which images belong to the same pedestrian. However, the annotation work faces many difficulties. On the one hand, the annotation process requires a large amount of manpower and time, resulting in low efficiency. On the other hand, due to modal differences, the appearance of

pedestrians in images of different modalities varies greatly, and manual annotation is prone to errors. Inaccurate annotation data will mislead the model training, causing the model to learn incorrect information and thus reducing the recognition performance of the model.

3.3 Application

Cross-modal person re-identification has extensive applications and great significance in multiple fields. In the field of security monitoring, it can achieve all-weather tracking of pedestrians through different modal image information. When an abnormal event occurs, even in the case of night or insufficient lighting, it can quickly lock the movement trajectory of the target pedestrian with the help of different modal data, enhancing the security monitoring ability and ensuring the safety and order of public places. In the intelligent transportation field, it can combine the RGB images of road surveillance cameras and the depth images obtained by depth sensors to analyze pedestrian flow and behavior patterns, count the number of pedestrians and recognize their actions, such as walking, running, and staying. These data provide a basis for urban traffic planning, help optimize traffic signal durations and rationally set up road facilities, improving the intelligent level of traffic management and alleviating traffic congestion. In the public safety field, it can be used for tracking criminal suspects. By using cross-modal data in different surveillance scenarios to match the identities of suspects and locate their positions, it can save the time and manpower costs of the police in solving cases. Moreover, in the security work of major events, it can screen key persons of concern in advance and monitor their activity paths in real time to prevent potential safety threats and ensure public safety.

4. DEEP REPRESENTATION LEARNING IN CROSS-MODALITY PERSON RE-IDENTIFICATION

4.1 Feature Extraction and Fusion

Feature Extraction Based on Convolutional Neural Network (CNN): CNN extracts the features of image modalities through different convolutional kernels. When processing RGB images, it captures visual features such as color and texture; when processing infrared images, it learns the features of the thermal radiation contour. The process is shown in Figure 3. Through multiple layers of convolution and pooling operations, the original pixels are converted into low-dimensional feature vectors, retaining the key information of pedestrian identity.

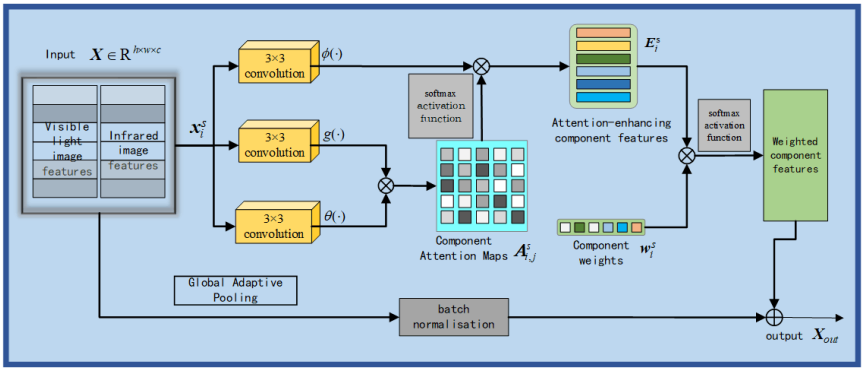

Figure 3: Convolutional neural network based feature extraction

Applications of Recurrent Neural Network (RNN) and Its Variants in Feature Extraction of Sequential Data: For sequential data such as videos, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) capture the action correlations between frames through memory units, and extract feature vectors containing temporal dynamic features. These feature vectors are combined with image features to enhance the cross-modal recognition ability.

Feature Fusion Strategies: Early Fusion: At the initial stage of feature extraction, multi-modal features (such as the concatenation of RGB and infrared features) are directly concatenated, and the concatenated features are provided for subsequent networks to jointly learn the correlations between modalities. Late Fusion: Single-modal models are trained separately, and at the decision-making layer, the results are fused through a weighted average or voting mechanism, retaining the advantages of the independence of each modality. Attention Fusion: Dynamically calculate the weights of modal features, adaptively allocate the contribution degrees of features, and focus on the more discriminative feature dimensions.

4.2 Model Building and Training

1) Design of the Multimodal Input Layer: Design the corresponding input layer according to the involved modalities (such as RGB images, infrared images, etc.). For example, for the image modality, the input layer needs to be adapted to the size, number of channels, etc. of the images. Input different modal data into their respective independent sub-network branches so that the network can extract features according to the characteristics of each modality.

2) Selection of the Feature Extraction Network: Select an appropriate feature extraction network for each modality. When dealing with the image modality, a Convolutional Neural Network (CNN) is often used. Classic CNN architectures such as ResNet and VGG can effectively extract features such as textures and shapes of images. If the video modality is involved, 3D-CNN can be used to capture spatio-temporal features. For the text modality (such as the textual information describing pedestrians), Recurrent Neural Networks (RNN) and their variants, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), can be employed to process sequential information.

3) Construction of the Modality Fusion Layer: After extracting the features of each modality, a fusion layer needs to be constructed for feature fusion. In early fusion, the features of each modality are directly concatenated at the initial stage of feature extraction, and then input into the subsequent network layers for joint training. In late fusion, after the feature extraction and classifier training of each modality are completed separately, the classification results are then fused. A fusion method based on the attention mechanism can also be adopted. By learning the weights of features of different modalities, the importance of each modality during fusion is dynamically allocated.

4) Setting of the Output Layer: The output layer usually adopts a fully connected layer. Determine the number of output nodes according to the task requirements (such as binary classification to judge whether it is the same pedestrian, or multi-class classification to identify the specific pedestrian identity), and select an appropriate activation function. If it is a classification task, the softmax function is commonly used to output the classification probabilities.

4.3 Ingoptimization Strategy

Selection of the Loss Function: The commonly used loss function is the cross-entropy loss function, which is used to measure the difference between the model's prediction results and the true labels and is suitable for classification tasks. In cross-modal person re-identification, to enhance the distinctiveness of features, the triplet loss function is often introduced. By defining the anchor sample, positive sample, and negative sample, this function makes the features of different modalities of the same pedestrian (positive sample and anchor sample) closer in distance, and the features of different pedestrians (negative sample and anchor sample) farther apart, thus optimizing the feature space and improving the recognition performance.

Selection of the Optimizer: Common optimizers include Stochastic Gradient Descent (SGD) and its variants such as Adagrad, Adadelta, and Adam. The Adam optimizer combines the advantages of Adagrad and RMSProp and can adaptively adjust the learning rate. It has a relatively fast learning speed in the early stage of training and can avoid the oscillation caused by a too large learning rate in the later stage, which is widely used in the training of cross-modal person re-identification models.

Data Augmentation: Since it is difficult to obtain cross-modal data, the amount of data is often limited. Data augmentation techniques can expand the scale of the data and improve the generalization ability of the model. For the image modality, operations such as rotation, translation, scaling, cropping, and adding noise can be performed; for the text modality, operations such as synonym replacement, random deletion, or insertion of words can be carried out.

Model Regularization: To prevent the model from overfitting, L1 and L2 regularization are often used. By constraining the weight parameters, the model's weights will not be too large, improving the generalization performance of the model. The Dropout technique can also be used. During the training process, some neurons are randomly discarded to avoid excessive dependence among neurons and reduce the risk of overfitting.

5. ISSUES AND CHALLENGES

The cross-modal person re-identification model has many problems: Firstly, the model has high complexity. The deep network constructs multiple layers of convolutional, recurrent structures and fusion modules to process multi-modal data, resulting in a large number of parameters and a complex structure, which makes training convergence difficult and the interpretability poor. Secondly, it has weak generalization ability, being sensitive to the data distribution, having poor adaptability across different scenarios (changes in lighting, posture, and background), and the insufficient coverage of training data affects the actual deployment effect. Thirdly, it consumes a large amount of computing resources. The training of a large number of parameters relies on high-performance GPUs, and a single training cycle is often long, making it difficult to adapt to scenarios with limited resources such as embedded devices. Fourthly, modal alignment is difficult. The heterogeneity of the cross-modal feature space makes it difficult for existing alignment methods to fully model the correlations between modalities, affecting the matching accuracy. Fifthly, there is a lack of high-quality annotated data. The high cost and low fault tolerance rate of multi-modal annotation lead to insufficient feature learning of the model, restricting the potential for performance improvement.

6. FUTURE DEVELOPMENT TRENDS

In the future, cross-modal person re-identification technology can make progress through a variety of methods: Data augmentation can be carried out by using Generative Adversarial Networks (GANs) to generate cross-modal samples under various postures and lighting conditions, alleviating the problem of data scarcity and improving the model's robustness to feature changes; Domain adaptation can be achieved with the help of transfer learning by migrating the pre-trained model from the source domain to the target domain and fine-tuning it to adapt to the features of the new environment, reducing the dependence on cross-scenario data; Privacy protection can be realized by applying federated learning. Under the distributed training mechanism, each institution only shares the gradient parameters of the model instead of the original multi-modal data, enabling multi-party collaborative modeling while ensuring privacy and security; Semantic fusion can be achieved through knowledge graphs by integrating pedestrian attributes and relationship graphs, enhancing the discriminative ability of features and reducing the false detection rate in complex scenarios; Lightweight model deployment can be realized by adopting technologies such as depthwise separable convolution and network pruning to compress the scale of model parameters (such as variants of MobileNet), meeting the real-time inference requirements of embedded devices.

7. CONCLUSION

The integrated research of deep representation learning and cross-modal person re-identification technology has achieved remarkable progress: The former has realized the efficient extraction of cross-modal features through multi-layer neural network architectures, while the latter has demonstrated important application value in fields such as security monitoring and intelligent transportation. The collaborative innovation of the two has effectively improved the accuracy of cross-scene pedestrian matching, yet it still faces core bottlenecks such as high model complexity and insufficient cross-scene generalization ability. This technological integration holds strategic significance for optimizing the public safety prevention and control system and the transformation towards smart cities. Future research will focus on key technological breakthroughs such as optimizing data distribution with generative adversarial networks, enhancing cross-domain adaptability through transfer learning, ensuring privacy and security with federated learning, strengthening semantic associations with knowledge graphs, and compatching model compression, continuously promoting the industrial implementation and ecological construction of multimodal artificial intelligence technology.

REFERENCES

[1] Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." IEEE transactions on pattern analysis and machine intelligence 35.8 (2013): 1798-1828.
[2] Zhong, Guoqiang, et al. "An overview on data representation learning: From traditional feature learning to recent deep learning." The Journal of Finance and Data Science 2.4 (2016): 265-278.
[3] Ju, Wei, et al. "A comprehensive survey on deep graph representation learning." Neural Networks (2024): 106207.
[4] Chen, Fenxiao, et al. "Graph representation learning: a survey." APSIPA Transactions on Signal and Information Processing 9 (2020): e15.

[5] Prajwal, Thode Sai, and Ilavarasi AK. "A comparative study Of RESNET-pretrained models for computer vision." Proceedings of the 2023 Fifteenth International Conference on Contemporary Computing. 2023.
[6] Jiang, Jianguo, et al. "A cross-modal multi-granularity attention network for RGB-IR person re-identification." Neurocomputing 406 (2020): 59-67.
[7] Hafner, Frank M., et al. "RGB-depth cross-modal person re-identification." 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2019.
[8] Zeng, Xuanli, et al. "Random area pixel variation and random area transform for visible-infrared cross-modal pedestrian re-identification." Expert Systems with Applications 215 (2023): 119307.
[9] Wu, Ancong, et al. "RGB-infrared cross-modality person re-identification." Proceedings of the IEEE international conference on computer vision. 2017.
[10] Wang, Yue. "Cross-Modality Person Re-Identification: An Attention-Enhanced Framework for Deep Fusion of Visible and Infrared Features." 2024 IEEE 16th International Conference on Computational Intelligence and Communication Networks (CICN). IEEE, 2024.
[11] Wang, Weidong, et al. "Feature decoupling and interaction network for defending against adversarial examples." Image and Vision Computing 144 (2024): 104931.
[12] Chan, Sixian, et al. "Parameter sharing and multi-granularity feature learning for cross-modality person re-identification." Complex & Intelligent Systems 10.1 (2024): 949-962.
[13] Zhang, Jiawei, et al. "Amc-net: An effective network for automatic modulation classification." ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023.
[14] Guo, Wenzhong, Jianwen Wang, and Shi** Wang. "Deep multimodal representation learning: A survey." Ieee Access 7 (2019): 63373-63394.
[15] Kazemi, Seyed Mehran, et al. "Representation learning for dynamic graphs: A survey." Journal of Machine Learning Research 21.70 (2020): 1-73.
[16] An infrared and visible image fusion algorithm based on ResNet-152
[17] Wang, Limin, et al. "Places205-vggnet models for scene recognition." arxiv preprint arxiv:1508.01667 (2015).
[18] Yang, Zihan. "Classification of picture art style based on VGGNET." Journal of Physics: Conference Series. Vol. 1774. No. 1. IOP Publishing, 2021.
[19] Zhou, Kaiyang, et al. "Omni-scale feature learning for person re-identification." Proceedings of the IEEE/CVF international conference on computer vision. 2019.
[20] Fukushima K. Neocognitron: A Self-Organizing Neural Network Model for A Mechanism of Pattern Recognition Unaffected by Shift in Position[J]. Biological cybernetics, 1980, 36(4): 193-202.
[21] Siami-Namini, Sima, Neda Tavakoli, and Akbar Siami Namin. "The performance of LSTM and BiLSTM in forecasting time series." 2019 IEEE International conference on big data (Big Data). IEEE, 2019.
[22] Zhang, Yaquan, et al. "Memory-gated recurrent networks." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 12. 2021.