IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 41, No. 1, February, 2024, pp. 189-199
Journal homepage: http://iieta.org/journals/ts

Multimodel-Based Gait Recognition Method with Joint Motion Constraints

Yanjun Qi*, Yi Xu, Xuan He

Institute of Information Security, Northwest University of Political Science and Law, Xi'an 710122, China

Corresponding Author Email: qiyanjun1@nwupl.edu.cn

https://doi.org/10.18280/ts.410115

Received: 15 August 2023
Revised: 2 December 2023
Accepted: 16 December 2023
Available online: 29 February 2024

Keywords:
constraint attention, gait recognition, motion constraint, multimodal

ABSTRACT

The intricate and restricted movements of joints form the core of pedestrian gait characteristics, with these traits externally reflected through the overall and synchronized gait movements. Thus, identifying features of such coordinated motions significantly boosts the discriminative effectiveness of gait analysis. Addressing this, we have introduced a novel gait feature mining approach that amalgamates multi-semantic information, effectively utilizing the combined strengths of silhouette and skeleton data through a meticulously designed dual-branch network. This network aims to isolate coordinated constraint features from these distinct modalities. To derive the coordinated constraint features from silhouette data, we crafted a silhouette posture graph, which employs 2D skeleton data to navigate through the silhouette's obscured portions, alongside a specialized local micro-motion constraint module. This module's integration of feature maps allows for the detailed extraction of features indicative of limb coordination. Concurrently, for the nuanced extraction of joint motion constraints, we developed a global motion graph convolution operator. This operator layers the motion constraint relations of physically separate joints onto the human skeleton graph's adjacency matrix, facilitating a comprehensive capture of both local and overarching limb motion constraints. Furthermore, a constraint attention module has been innovated to dynamically emphasize significant coordinated motions within the feature channels, thus enriching the representation of pivotal coordinated motions. This advanced network underwent thorough training and validation on the CASIA-B dataset. The ensuing experimental outcomes affirm the method's efficacy, demonstrating commendable recognition accuracy and remarkable stability across varying viewing angles and dynamic walking conditions.

1. INTRODUCTION

Compared to traditional biometric recognition technologies, gait recognition offers unparalleled advantages, primarily due to its independence from subject cooperation and effective applicability in long-range scenarios. This technique has extensive applications across various domains, including public safety, medical research [1, 2], and crime prevention [3, 4]. The variability in walking conditions and environmental factors, combined with the inherently nonrigid nature of pedestrian motion, tends to amplify the interclass distance of the same individual, thereby affecting gait recognition performance. Gait, a complex coordination of bodily joints and bones, is intricately defined by the interactions between joints (such as elbows and knees) and the specific movements of limbs (including knees and hip joints). The extraction of these relational features plays a crucial role in identifying unique and distinctive gait characteristics, thereby enhancing the precision of gait analysis.

To acquire robust pedestrian gait features, researchers have developed sophisticated models that provide an abstract representation of gait. These models capitalize on the advanced nonlinear feature modeling capabilities of deep learning and extract identifiable characteristics crucial for gait recognition. Currently, the most effective methods are based on silhouette-based methods [5-8] and skeleton-based methods [9-11]. Both strategies integrate dynamic attributes of gait, concentrating on temporal variations within comprehensive spatial information and thereby substantially enhancing the precision of gait recognition. Silhouette-based methods extract binary images of human contours from video data and craft convolutional neural network (CNN) models that are proficient in delineating the intricate spatiotemporal characteristics of these contours. Notable examples include GaitSet [5], which aggressively discerns the positional interrelationships of unordered contours, and GaitPart [12], which concentrates on the movement dynamics of specific body segments. These approaches, by attentively focusing on spatial variances in human form, have achieved state-of-the-art performance in gait recognition. Nevertheless, gait silhouette images, despite effectively conveying spatial motion information, are susceptible to environmental influences in dynamic settings and fail to capture spatiotemporal information about joint constraints during limb self-occlusion [13]. With the advancements in human pose estimation methods [14, 15], researchers have been able to directly acquire joint coordinates from videos, efficiently encoding these coordinates into human skeleton graphs. By

utilizing the node feature aggregation capability of graph convolution networks (GCNs), these methods capture the spatiotemporal motion features of gait. For instance, the ResGCN [9] aggregates features of physical structure-connected joints using a GCN in the spatial domain and employs CNNs for temporal aggregation of joint movement. These innovative methods greatly mitigate the impact of various impediments, such as physical obstructions, lighting discrepancies, and viewpoint variations, on gait recognition accuracy. However, skeletal data, despite their informative nature, do not fully encompass the external characteristics of the human form.

Considering that silhouette data provide rich spatial post information and skeleton data captures joint motion dynamics, researchers have strategically combined these two modal datasets to harness their complementary advantages. For example, Wang and Chen [16] employed a dual-branch network to explore feature mining methods across these modalities. This method constructs a fully connected graph to aggregate features of non-physically connected joints. However, such full connectivity can weaken the features of coordinated movements in key joints. Nonetheless, there is a need for further research into the extraction of coordinated limb movement features from silhouette images. Gait, inherently, is an orchestration of the entire body's movement. It includes the linked movements of adjacent joints, like elbows and wrists, to articulate local human motion, as well as the interactions between physically disconnected joints, such as elbows and knees, to represent global gait movement. This comprehensive movement mirrors the coordination and coherence of human motion. The identification and extraction of these intricate relationships are crucial in augmenting the discriminatory power of gait features.

This study delves into the articulation of motion constraint relationships between limbs across various modalities, introducing a gait constraint feature mining method that amalgamates multisemantic information. This approach highlights the synchronicity of motion across the human body, both globally and locally, by formulating a global motion adjacency matrix that delineates the coordinated movements of physically disconnected joints. The core contributions of this paper are outlined as follows:

(1) To distill synergistic motion features from contour data, we devised contour posture maps, employing skeletal data to steer the limb movements within the contour. A local micro-motion constraint module was crafted to synthesize the movements of body segments in a sequential manner, capturing the collaborative motion features of two parts over a specified duration.

(2) A constrained motion graph convolutional operator was designed, capturing both the local motion of joints and the collaborative movement of distal joints within an adjacency matrix. This operator adeptly extracts the local dependency features of joints alongside the global dependency features of distal joints, ensuring a comprehensive representation of motion.

(3) We established a dual-branch gait constraint feature mining network, which leverages attention mechanisms to amplify the key synergistic motion features. This network achieved commendable recognition rates on the CASIA-B database, showcasing enhanced recognition stability across wide-ranging viewpoints and variable walking conditions.

2. RELATED WORKS

Gait recognition methods can be broadly categorized into appearance-based methods [5-7, 17-19] and model-based methods [9-11, 20, 21], differentiated by the underlying gait feature description models employed. Both the silhouette-based method of the former and the skeleton-based methods of the latter achieve favorable recognition results because of the focus on temporal variations in gait.

Silhouette-based methods first extract pedestrian silhouettes from videos to construct a gait silhouette graph. It fully expresses the motion characteristics of the human silhouette in two dimensions. Researchers such as Thapar et al. [22], Li et al. [23], and Wolf et al. [24] have effectively addressed the extraction of spatiotemporal gait features by employing long short-term memory (LSTM) networks and 3D-CNNs. Furthermore, Chao et al. [5] attempted to input unordered sets of gait silhouette graphs into the GaitSet network, automatically learning the spatial motion and positional relationships of gait. Fan et al. [12] focused on the motion characteristics of different body parts, designed the GaitPart network and used the micromotion capture module (MCM) to model local micromotion features, thereby obtaining spatiotemporal motion features of different parts. Hou et al. [25] constructed the gait lateral network (GLN) to learn distinctive, compact feature representations from silhouettes. Then, they merged features extracted at different stages at the silhouette level and set level using a feature pyramid in a top-down and lateral connection fusion approach. Although these methods can model the spatial information of humans in the temporal dimension, in complex scenes, gait silhouette graphs might introduce irrelevant information. Moreover, when the body self-occludes, the motion information between limbs is weakened, limiting further enhancement of the performance of these methods. Interventionary studies involving animals or humans and other studies that require ethical approval must list the authority that provided approval and the corresponding ethical approval code.

Skeleton-based methods employ human pose estimators to directly estimate two-dimensional or three-dimensional joint coordinates from images or videos, effectively minimizing external environmental influences. For example, Liao et al. [20] and Qi et al. [21] innovatively encoded interlimb motion relationships into pseudoimages and developed networks such as PoseGait and LC-POSEGAIT for gait feature modeling. However, given the holistic nature of human motion, studies often fail to fully capture the intricate motion dependencies and constraints among limbs. In recent years, GCNs have emerged as potent tools for feature modeling in non-Euclidean space data. As an effective feature extractor for graph-structured data, a GCN excels in aggregating features from neighboring nodes, thus enabling efficient feature transmission. Scholars have built GCN-based networks to directly feature-model human skeleton graphs. These methods lead to substantial improvements in recognition rates over methods that extract features from manually crafted pseudoimages. Yan et al. [26] initially designed the ST-GCN for feature modeling of human skeleton graphs. Inspired by these previous works, Teepe et al. [9] proposed the ResGCN network. This network, comprising each residual block of the GCN and CNN, adeptly aggregates joint features across both spatial and temporal dimensions. To solve the problem of weight bias due to the aggregation of long-range joint features, Hasan et al. [27] proposed a GCN with multiscale feature

aggregation technology. This approach involves creating adjacency matrices for varying distance hops, effectively reducing bias in feature aggregation.

Utilizing different modal data, such as deep sensors and videos, for gait recognition can enhance recognition performance [28]. Li et al. [29] combined the advantages of silhouettes and posture heatmaps to construct ensemble transformer modules, modeling motion patterns across different time scales. Wang and Chen [16] constructed a dual-branch neural network model for feature-model silhouette and skeleton data. Peng et al. [30] proposed the BiFusion network to mine the complementary cues of skeletons and contours. This method constructs skeletal graphs at three scales-joints, limbs, and torso-based on the inherent hierarchical semantics of human joints within the skeleton.

Although scholars have explored multimodal gait feature mining methods, further strengthening the expression of gait features through holistic human motion and interactive constraint features between limbs is possible. Therefore, this study combines the expressive advantages of multimodal gait data to mine constraint features between joints, aiming to obtain a discriminative gait feature representation.

3. JOINT MOTION CONSTRAINTS FEATURE EXTRACTION METHODS

To obtain the motion information of obscured parts in gait silhouette graphs, we leverage the advantage of skeleton graphs by superimposing two-dimensional human posts onto silhouette images, thus achieving a comprehensive representation of limb motion and shape information.

3.1 Multimodal data construction

To obtain the motion information of obscured parts in gait silhouette graphs, we leverage the advantage of skeleton graphs by superimposing two-dimensional human posts onto silhouette images, thus achieving a comprehensive representation of limb motion and shape information. As shown in Figure 1, the left side shows the video sequence of subject 004-bg-01-054 from the CASIA-B dataset, the middle side shows the gait silhouette and 2D skeleton graphs, and the right side shows the combined silhouette post graph G_{sj}. It is evident from the figure that the spatial motion information of the occluded right arm of the pedestrian in motion can be reproduced using the skeleton graph.

For the human skeleton graph, the sequence of human skeleton graphs obtained from three-dimensional post estimation clearly expresses the spatial position changes of joints. Additionally, pedestrian motion speed and body sway are also typical features of gait. Compared to single-semantic gait description data, multiangle feature descriptions can more richly express gait characteristics in three-dimensional space. Here, the motion speed and body sway characteristics are abstractly described based on the three-dimensional coordinates of the joints. The human skeleton graph is denoted as, where the point set $V = \{v_1, v_2, \cdots, v_N\}$ represents N joint nodes, and the edge set $E = E_s \cup E_t$ consists of spatial edges E_s and temporal edges E_t . $E_s = \{(v_i, v_j)|\ v_i, v_j \in V\}$ represents the edges formed by bone connections between joints v_i and v_j, and $E_t = \{v_i^t, v_i^{t+1})|t, t+1 \in T\}$ represents the temporal edge formed by joint v_i between frames t and $t+1$.

The joint coordinate data of the skeleton graph are denoted as D_j. The pedestrian's walking speed is described by the difference in joint coordinates between two frames, denoted as D_d , as shown in Eq. (1), where a^t, b^t, c^t , respectively represent the displacements of joint v_i in the x, y, z directions between adjacent frames.

$$D_d = \{(a^t, b^t, c^t)_i|i \in N, t \in T\} \tag{1}$$

The degree of body sway is represented by the angle θ formed by the vector l_{Gi}, which is composed of joint v_i and the center of gravity v_G, with the normal vector n of the coordinate plane denoted as D_a, as shown in Eq. (2), where $\theta_x^t, \theta_y^t, \theta_z^t$ represents the angles between vector l_{Gi} and the three planes.

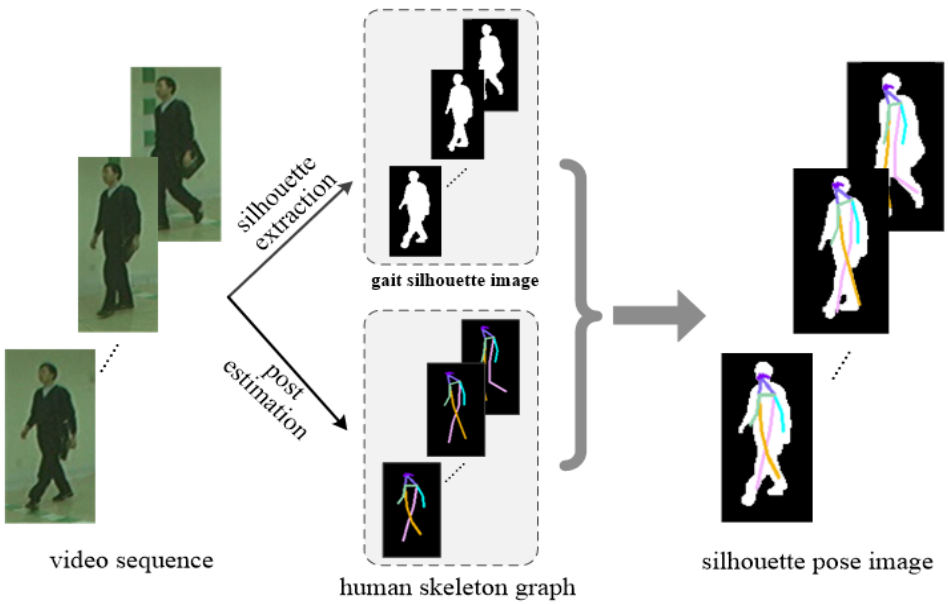

Figure 1. Silhouette, skeleton, and silhouette post graphs of subject #004 (bg-01-054) in the CASIA-B dataset

$$\begin{cases} D_a = \{(\theta_x^t, \theta_y^t, \theta_z^t) | i, j \in N, t \in T\} \\ \theta^t = \cos^{-1}\left(\dfrac{\overline{l_{Gi}^t} \cdot \bar{n}}{|\overline{l_{Gi}^t}|\,|\bar{n}|}\right) \end{cases} \tag{2}$$

A multisemantic dataset of joint motion relationships $D = \{D_j, D_d, D_a\}$ is constructed, resulting in the multimodal gait dataset $D_{gait} = G_{sj} \cup D$.

3.2 Gait constrained feature mining network

To obtain unique abstract gait features from diverse semantic gait data, a dual-branch network is used to model features of two types of data according to their specific characteristics. The silhouette post graph G_{sj}, as grid data with translational invariance, enables the convolutional kernels of CNNs to focus on local changes in the image. This approach effectively captures the features of crucial areas, harnessing the capacity of CNNs to model local features and track motion changes in humans. In contrast, the human skeleton graph G represents non-Euclidean space data and lacks translational invariance. A GCN synthesizes features from a node and its adjacent nodes, thereby generating novel features for the node and capturing the global information of the graph-structured data. Owing to the significant advantage of GCN networks in extracting features from graph-structured data, they are the preferred choice for modeling features of the human skeleton.

Gait embodies a complex coordination of joints, bones, and muscle tissues, forming an integrated movement pattern. The identification of interactive constraint features between limbs and joints is crucial for improving feature distinction. Temporally, the significance of a particular body part or joint's movement fluctuates over different periods. Spatially, the interactions among joints vary due to movement constraints. Consequently, a novel approach involves designing a constraint attention module. This module implicitly assesses the movement significance of limbs and joints, thereby extracting significant features effectively.

This study introduces the gait-constrained feature extraction network based on attention (Att-CFEN), as shown in Figure 2. The network comprises three primary components: the Limb Constraint Feature Extraction sub-Network (LCFEN), the Joint Constraint Feature Extraction sub-Network (JCFEN), and the Fusion Attention Module (FAM). The LCFEN and JCFEN are responsible for processing the silhouette posture graph and the human skeleton graph, respectively. These methods produce advanced limb constraint feature vectors Y_A and joint constraint feature vectors Y_B. Subsequently, the FAM integrates these feature vectors from both modalities through weighted fusion. This process evaluates the relative importance of different modal feature vectors, leading to gait classification outcomes via a fully connected layer. The feature modeling procedures for both branches are described in the following sections.

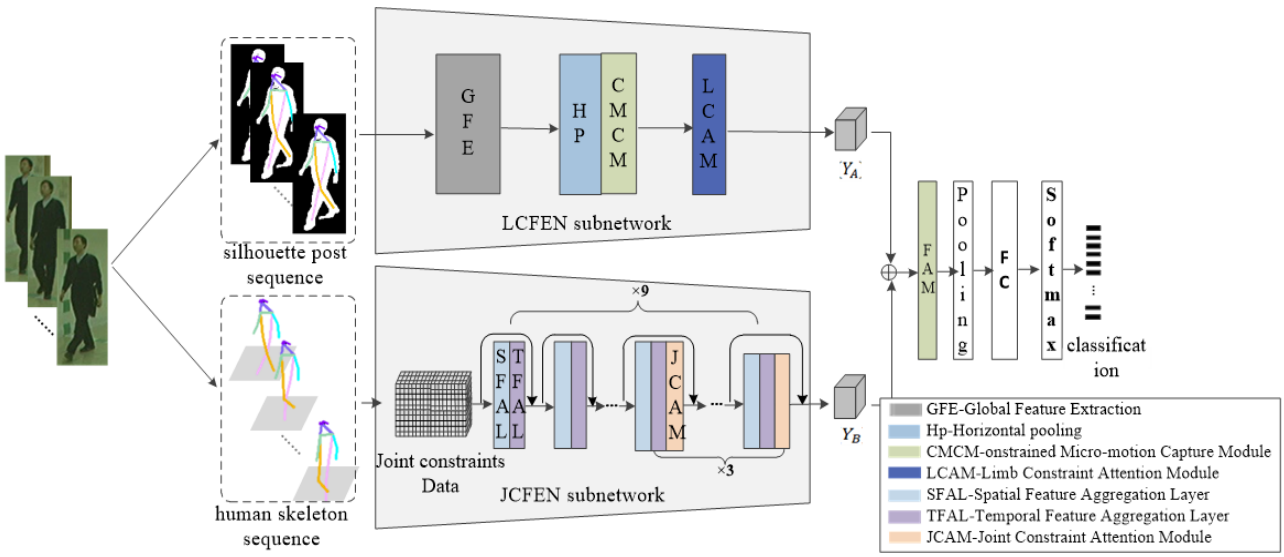

Figure 2. Att-CFEN network framework diagram

3.2.1 Limb constraint feature extraction

For the silhouette post graph, the analysis extends beyond global motion information to include constrained movements between limbs (e.g., arms, legs, and combinations thereof), which are indicative of individual walking patterns. The Gaitpart model [12] demonstrates exceptional capability in capturing micro movements between limbs. Consequently, the LCFEN architecture, as shown in Figure 3, is grounded in the Gaitpart framework for effective feature modeling of these limb constraint micro movements.

The input to the LCFEN is a contour posture map, whose features $X_A^{(0)} \in \mathbb{R}^{C \times H \times W}$ are a three-dimensional tensor with dimensions $C \times H \times W$, where C represents the feature dimension, and H and W are the height and width of the feature map, respectively. The Global Feature Extractor (GFE) employs focal convolution for fine-grained spatial feature extraction, obtaining frame-level local spatial features X_A. Horizontal Pooling (HP) first divides $X_{A(F)}$ horizontally into n parts, denoted as $[X_{A(1)}, X_{A(2)}, \cdots, X_{A(n)}]$, where $X_{A(i)} \in \mathbb{R}^{\frac{C}{n} \times H \times W}$, segmenting the body into n local regions. Besides extracting frame-level features for each part, it also performs frame-level feature fusion, combining the frame-level features of body parts to obtain fused part features. Then, it fuses part features with frame-level features to achieve limb synergistic motion features.

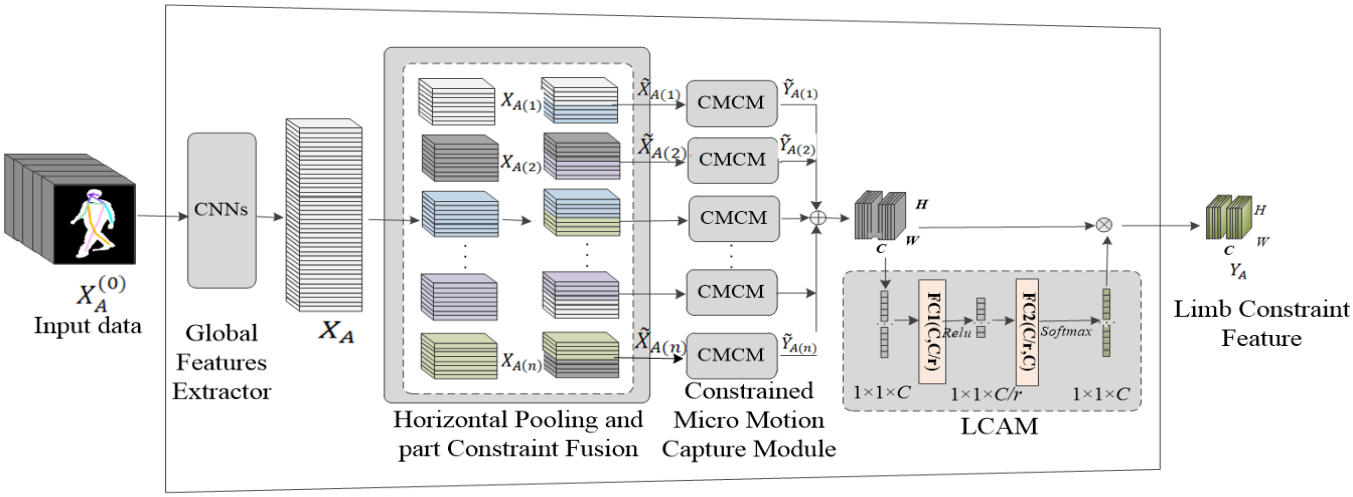

Figure 3. Structure of the LCFEN

Specifically, the method merges the motion frames of the front and back halves of two body parts at a step size. Let the fusion step size be m, meaning the i-th part is fused with the $(i+m)$-th part; the fusion scale is set to 2, indicating that the second half of the feature map of the i-th part swaps places with the second half of the feature map of the $(i+m)$-th part, as shown in Eq. (3). The fused feature maps of n parts are denoted as $[\tilde{X}_{A(1)}, \tilde{X}_{A(2)}, \cdots, \tilde{X}_{A(n)}]$.

$$\tilde{X}_{A(i)} = [x_{A(i)}]_j \oplus [x_{A(i+m)}]_k (j \in (1, \frac{c}{2n}), k \in (\frac{c}{2n}+1, c)) \tag{3}$$

The constrained micromotion capture module (CMCM) transforms the fused features $[\tilde{X}_{A(1)}, \tilde{X}_{A(2)}, \cdots, \tilde{X}_{A(n)}]$ into limb constraint features $[\tilde{Y}_{A(1)}, \tilde{Y}_{A(2)}, \cdots, \tilde{Y}_{A(n)}]$. Subsequently, these features are integrated into a comprehensive global constraint feature $\tilde{Y}_A$, as detailed in Eq. (4).

$$\tilde{Y}_A = Concat(\tilde{Y}_{A(1)}, \tilde{Y}_{A(2)}, \cdots, \tilde{Y}_{A(n)}) \tag{4}$$

To obtain significant constraint features of body parts, the limb constraint attention module (LCAM) is integrated into the system. The LCAM evaluates constraint features across various channels, emphasizing key gait characteristics and thereby augmenting the distinctiveness of gait features. Utilizing channel attention, LCAM assesses the feature maps. Its structural design is illustrated within the dashed box in Figure 4. The LCAM executes a one-dimensional convolution on the feature channels. This process begins with a global average pooling that condenses $\tilde{Y}_A$ into a $1 \times 1 \times C$ vector, as shown in Eq. (5).

$$f_1(\tilde{Y}_A) = \frac{1}{W \times H} \sum_{j=1}^{W} \sum_{k=1}^{H} \tilde{Y}_A(j,k) \tag{5}$$

The vector resulting from the LCAM is then processed through two fully connected layers, FC1 and FC2. The first layer, FC1, reduces the dimensionality from C channels to C/r channels, enhancing computational efficiency, where r is a predefined positive integer. FC2 then maps this reduced dimensionality back to C channels. Following this, the Softmax function is applied to derive the weight W_t for each channel. These weights are then elementwise multiplied by $\tilde{Y}_A$ to produce the weighted feature map Y_A, as depicted in Eq. (6). Consequently, the HCFEN ultimately outputs a feature set Y_A that accurately represents the coordinated movements of the human body's limbs.

$$\begin{cases} W_l = Softmax(FC2\left(FC1\left(f_1(\tilde{Y}_A)\right)\right)) \\ Y_A = W_t \otimes \tilde{Y}_A \end{cases} \tag{6}$$

3.2.2 Joint constraint feature extraction

For physically connected joints, such as the shoulder and elbow joints, they can express the local motion of limbs, manifesting as constrained motion relationships between joints. Conversely, physically unconnected joints, such as the elbow and knee joints, can reflect the coordination of gait, exhibiting synergistic motion relationships between them. In traditional Graph Convolutional Networks (GCNs), node features are aggregated based on the encoding of the adjacency matrix, which typically encodes only the nodes that are connected. This approach can lead to the attenuation of features for distal joints due to the weighted bias in the conventional graph convolution operators. Therefore, to aggregate the features of physically unconnected joints onto their interacting joints effectively, a new method encodes the constraint relationship between joints as $\tilde{a}_{i,j}$, according to Eq. (7), where E_u represents the constraint edges between physically unconnected joints. This method establishes connectivity edges between symmetrical limbs' joints, such as the elbow, wrist, knee, and ankle joints, directly expressing their constrained motion relationships.

$$\tilde{a}_{i,j} = \begin{cases} 1, & (v_i, v_j) \in E_s \cup E_u \\ 0, & (v_i, v_j) \notin E_s \cup E_u \end{cases} \tag{7}$$

All constraint relationships form the adjacency matrix, named the global motion adjacency matrix $\tilde{A} = [\tilde{a}_{i,j}] \in \mathbb{R}^{N \times M}$, where N and M represent the number of joints and spatial edges, respectively. The GCN processes spatial features of nodes based on the encoding of $\tilde{A}$, as detailed in Eq. (8). In this context, D is the degree matrix of $\tilde{A}$, $\hat{D}$ is the inverse matrix of the joint degree matrix D, $H^{(l)}$ denotes the feature of the l-th layer, σ is the activation function, and $W^{(l)}$ represents the parameter matrix of the l-th layer. The temporal feature

aggregation for joints is conducted according to Eq. (9), where f_i is the feature of the i-th joint and T signifies the aggregation time.

$$H^{(l+1)} = \sigma[\hat{D}^{-\frac{1}{2}}\tilde{A}\hat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}] \tag{8}$$

Inspired by the ST-GCN action recognition network, the Joint Constraint Feature Extraction subnetwork (JCFEN) is designed to model the constraint features of joints, as depicted in Figure 4. This network processes the multisemantic dataset D, constructed in Section 3, with its input features represented as $X_{B(1)}^{(0)}, X_{B(2)}^{(0)}, X_{B(3)}^{(0)} \in \mathbb{R}^{C\times T\times K}$. Here, C is the joint feature dimension, T is the number of frames, and K is the number of joints. JCFEN employs nine spatiotemporal feature aggregation modules (STFAMs) with residual connections for skeletal feature modeling. Each STFAM consists of a spatial feature aggregation layer (SFAL) and a temporal feature aggregation layer (TFAL). In parallel with the LCFEN subnetwork, the Joint Constraint Attention Module (JCAM) is employed following the last three aggregation modules. The JCAM is instrumental in obtaining feature maps that highlight the spatiotemporal significance of joint movements, thus enhancing the representation of high-level semantic joint features.

Within this subnetwork, three types of data are processed through six STFAMs, resulting in three sets of features $X_{B(1)}^{(i)}, X_{B(2)}^{(i)}, X_{B(3)}^{(i)}$. These feature sets are subsequently combined to produce the joint fusion feature $X_B^{(i)}$. Subsequently, the JCAM refines these features to generate the weighted feature $X_B^{(i+1)}$. The structure of the JCAM module is illustrated in Figure 5.

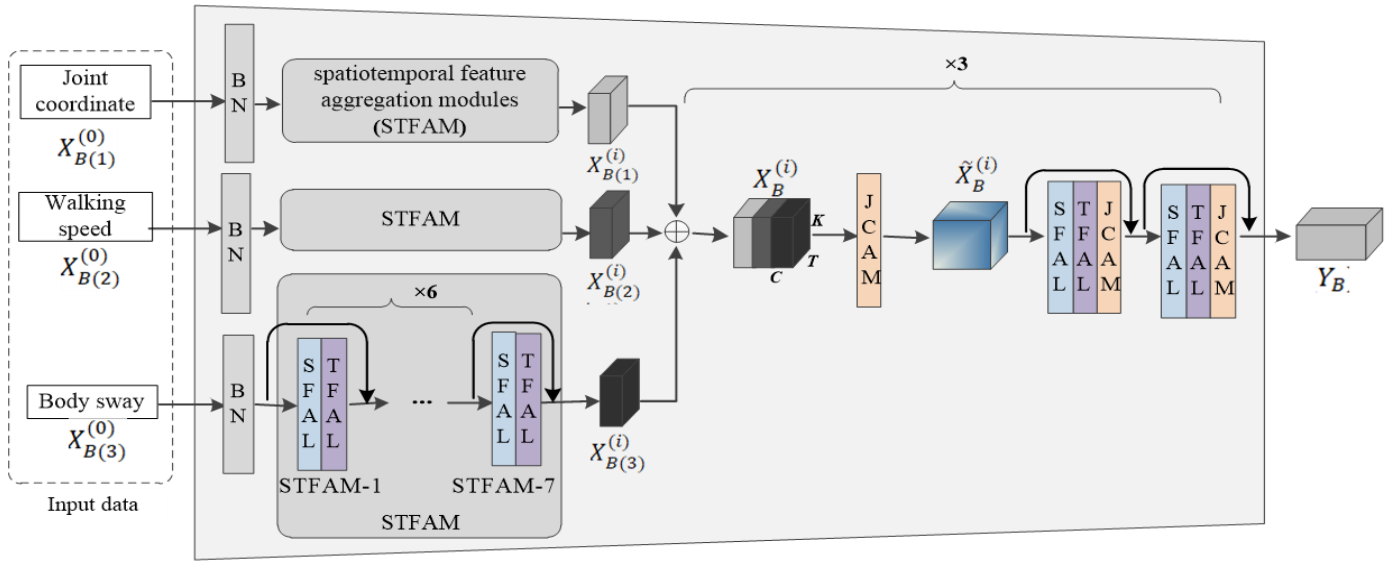

Figure 4. Structure of the JCFEN

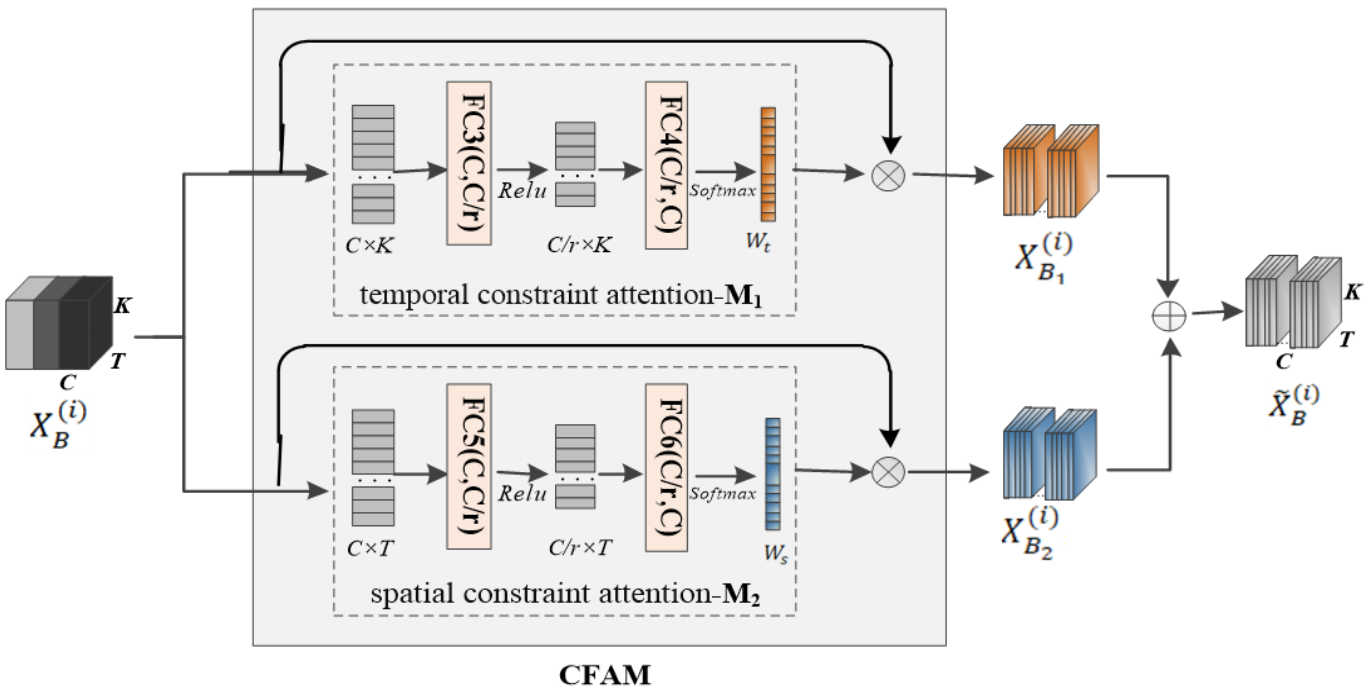

Figure 5. Structure of the JCAM

The JCAM in this network emphasizes the significance of joint movement from temporal and spatial dimensions. It utilizes temporal constraint attention M1 and spatial constraint attention M2 to apply importance weighting to the feature map. M1 focuses on the spatial motion features of the joints over time by compressing the input joint motion feature map along the temporal dimension. This process yields temporal feature scoring for each joint within the feature channels. Conversely, M2 focuses on the coordinated motion features between joints by compressing the feature map along the spatial dimension. It compresses the feature maps along the spatial dimension to obtain the feature scores of motion frames, ultimately generating the joint constraint feature map $\widetilde{\mathrm{X}}_{\mathrm{B}}^{(\mathrm{i})}$. The architectures of M1 and M2 align with the JCAM and are therefore not repeated here. The specific functions of these models are as follows.

M1 averages the joint features in $X_B^{(i)}$ along the temporal dimension, expressed as $\frac{1}{T}\sum_{t=1}^{T} X_B^{(i)}$, to capture the global context feature of each feature map. Subsequently, the fully connected layer FC4 remaps the tensor from FC3, from a dimension of $\mathrm{C/r} \times \mathrm{K}$ to $\mathrm{C} \times \mathrm{K}$. The Softmax function is then employed to determine the weight $\mathrm{W_t}$ of each joint. These weights are applied to $X_B^{(i)}$ to produce the feature map $\mathrm{X}_{\mathrm{B}_1}^{(\mathrm{i})} \in \mathrm{R}^{\mathrm{C}\times\mathrm{T}\times\mathrm{K}}$ with joint weights. The implementation of M1 is detailed in Eq. (9). Here, σ represents the Softmax activation function.

$$\begin{cases} W_t = \sigma(FC4(FC3(Avgpool(X_B^{(i)})) \cdot w_1) \cdot w_2) \\ X_{B_1}^{(i)} = X_B^{(i)} \otimes W_p \end{cases} \tag{9}$$

M2 processes the weighted feature map $\widetilde{\mathrm{X}}_{\mathrm{B}}^{(\mathrm{i})}$ by averaging along dimension K, specifically calculating $\frac{1}{\mathrm{K}}\sum_{\mathrm{k}=1}^{\mathrm{K}} \widetilde{\mathrm{X}}_{\mathrm{B}}^{(\mathrm{i})}$, to assign frame-level weight scores. These data are then traversed through the fully connected layers FC5 and FC6, resulting in the determination of frame-level weights $\mathrm{W_p}$. These weights are subsequently applied to $\tilde{X}_B^{(i)}$ to produce the final spatiotemporal weighted feature map $\mathrm{X}_{\mathrm{B}_2}^{(\mathrm{i})}$. The operational methodology of M2, encompassing this entire process, is shown in Eq. (10).

$$\begin{cases} W_P = \sigma(FC6(FC5(Avgpool(X_B^{(i)})) \cdot w_3) \cdot w_4) \\ X_{B_2}^{(i)} = X_B^{(i)} \otimes W_p \end{cases} \tag{10}$$

Then, $\mathrm{X}_{\mathrm{B}_1}^{(\mathrm{i})}$ and $\mathrm{X}_{\mathrm{B}_2}^{(\mathrm{i})}$ are combined through the $\oplus$ operation to obtain the spatiotemporal weighted feature map of joint constraints, $\widetilde{\mathrm{X}}_{\mathrm{B}}^{(\mathrm{i})}$, as shown in Eq. (11).

$$\tilde{X}_B^{(i)} = X_{B_1}^{(i)} \oplus X_{B_2}^{(i)} \tag{11}$$

3.2.3 Fusion feature modeling

Once the subnetworks LCFEN and JCFEN generate the feature vectors $\mathrm{Y_A} \in \mathbb{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$ and $\mathrm{Y_B} \in \mathbb{R}^{\mathrm{C}\times\mathrm{T}\times\mathrm{K}}$, respectively, representing constraints between limbs and joints, the subsequent phase of the Att-CFEN network (illustrated in Figure 3) employs the FAM. The role of the FAM is to evaluate the contributions of these two modal features. This evaluation involves pooling operations on both types of constraint vectors to condense the feature channel dimensions and assign scores to the resultant feature maps. The role of the FAM is to evaluate the contributions of these two modal features. This evaluation involves pooling operations on both types of constraint vectors to condense the feature channel dimensions and assign scores to the resultant feature maps. While the attention modules in the preceding branches use average pooling to capture the global features of the maps, the FAM incorporates both maximum pooling and average pooling. This dual approach enables the FAM to concentrate on features from varying perspectives: Maxpool targets the most significant features of the map, whereas Avgpool focuses on the global features. Following these pooling operations, a $1\times1\times2\mathrm{C}$ dimensional feature vector X' is created, as detailed in Eq. (12). This vector is then processed through two fully connected layers to produce a fusion weight matrix W, calculated similarly to the method outlined in Eq. (11), where W and X' yield a weighted feature map.

$$X' = \left(Maxpool(Y_A) \oplus Avgpool(Y_A)\right) \oplus (Maxpool(Y_B) \oplus Avgpool(Y_B)) \tag{12}$$

Thus, the FAM outputs the weighted feature map of the two modalities combined. After passing through the pooling layer and the fully connected layer, the dimension reduction and one-dimensional linear transformation of the fused weighted features are completed, and pedestrian feature classification is performed using the Softmax function.

4. EXPERIMENTS AND ANALYSIS

This study implemented network training on the publicly available gait dataset CASIA-B and devised three experiments to assess the efficacy of the Att-CFEN network. The first experiment involved a statistical analysis of the recognition rates achieved by the new method, and compared with benchmark techniques in pose estimation to evaluate its performance. In the second experiment, ablation studies were conducted to evaluate the contribution of the constraint attention module in accentuating significant features. Finally, the third experiment focused on examining the role of multisemantic data in enriching gait feature representation.

4.1 Experimental datasets

The CASIA-B multiview gait dataset contains 124 pedestrians, each recorded under three walking conditions: walking with a bag (BG), wearing a coat (CL), and walking normally (NM). These were recorded from 11 angles, ranging from 0° to 180°, yielding a total of $124\times10\times11$ video clips for each participant. The 2D coordinates of pedestrian joints were extracted from these videos using the HRNet algorithm and subsequently transformed into 3D coordinates. For network training, the dataset included all ten walking states from pedestrians #001 to #074. The gallery set utilized the nm01-04 data of pedestrians #075 to #124. Moreover, the probe set included NM05-06, BG01-02, and CL01-02 data from pedestrians #075 to #124, providing a comprehensive range of gait patterns for evaluation.

4.2 Experimental analysis

4.2.1 Recognition rate analysis

The recognition rates of the Att-CFEN network on the CASIA-B dataset were subjected to a thorough statistical analysis, as detailed in Table 1. This analysis computed the average recognition rates for the NM, BG, and CL walking states across 11 angles. These rates were subsequently contrasted with several established methods: PoseGait [20] for pose estimation, Gaitgraph [9] and GaitGraph2 [31] for graph convolution, GaitSet [5] for gait silhouette analysis, method [16] and BiFusion [30] for multimodal. The findings, as presented in the table, indicate that the average recognition rates of Att-CFEN under the three conditions were 93.0%, 87.2%, and 82.5%, respectively, surpassing those of the compared methods. The recognition rate under the CL condition was lower than that under the other two conditions, which was primarily attributed to the impact of a coat on the pedestrian's silhouette. This factor adversely affects the pose estimation accuracy, thereby diminishing the recognition performance. Compared to the PoseGait method, the Att-CFEN network employs a more precise HRnet for pose estimation and utilizes a GCN to discern motion constraint relationships between joints. This approach, coupled with the analysis of human form characteristics, leads to a significant improvement in recognition rates. In contrast to GaitGraph and GaitGraph2, the Att-CFEN method not only uses three-dimensional joint coordinates but also integrates additional features such as joint centroid offset angles and joint displacements. This approach enriches the expression of joint motion characteristics, enabling the network to model gait features from multiple perspectives. Additionally, the application of spatiotemporal attention mechanisms effectively highlights key motion features, amplifies crucial gait characteristics, and minimizes irrelevant factors, thereby markedly enhancing feature discrimination.

While GaitSet achieves higher recognition rates under NM conditions than does the Att-CFEN method, the latter outperforms GaitSet under CL conditions by a significant margin of 12.1 percentage points. For example, GaitSet's recognition rate at a 0° angle drops from 90.8% (NM) to 61.4% (CL), a substantial decrease of nearly 30 percentage points; a similar trend is observed at a 180° angle. However, the Att-CFEN demonstrated remarkable stability under various walking conditions (BG, CL), with only a 2.2% reduction in the average recognition rate. This demonstrates the robustness of the Att-CFEN in terms of recognition performance. The average recognition rate of Att-CFEN is slightly lower than that of methods presented in the study by Teepe et al. [9]. This discrepancy is attributed to the estimation of human skeleton joint coordinates, which initially involves two-dimensional posture estimation before transitioning to three-dimensional estimation, introducing potential data inaccuracies. Despite this, Att-CFEN maintains comparable recognition rates at extreme viewing angles (0° and 180°) relative to other angles. This consistency suggests that the incorporation of three-dimensional coordinates effectively mitigates the impact of varying shooting angles on gait recognition, demonstrating the method's adaptability to different observational perspectives.

Compared to the multimodal BiFusion [30] method, the method presented in this paper achieves an average recognition rate of 86.5%, which is lower than that of the BiFusion method. However, under the CL walking condition, the difference between the highest and lowest recognition rates from different viewpoints is 5.4% for our method, compared to 8.2% for the BiFusion method. The reason for this is that our method employs three-dimensional joint coordinates, which offer better viewpoint robustness. However, due to biases introduced by two stages of pose estimation, the average recognition rate of our method is lower than that of the BiFusion method.

Table 1. Average recognition rates (%) on the CASIA-B dataset

Probe		Gallery: NM01-04（#075-#124）											
		0º	18º	36º	54º	72º	90º	108º	126º	144º	162º	180º	Mean
NM 05-06	PoseGait [20]	55.3	69.6	73.9	75.0	68.0	68.2	71.1	72.9	76.1	70.4	55.4	68.7
	GaitGraph [9]	85.3	88.5	91.0	92.5	87.2	86.5	88.4	89.2	87.9	85.9	81.9	87.7
	GaitGraph2 [31]	78.5	82.9	85.8	85.6	83.1	81.5	84.3	83.2	84.2	81.6	71.8	82.0
	GaitSet [5]	90.8	97.9	99.4	96.9	93.6	91.7	95.0	97.8	98.9	96.8	85.8	95.0
	Multimodal [16]	97.0	97.9	98.4	98.3	97.2	97.3	98.2	98.4	98.3	98.1	96.0	97.7
	BiFusion [30]	98.0	99.1	99.5	99.3	98.7	97.5	98.5	99.1	99.6	99.5	96.8	98.7
	Att-CFEN	**92.1**	**92.5**	**93.2**	**93.8**	**92.7**	**92.9**	**94.3**	**94.4**	**93.7**	**93.6**	**90.1**	**93.0**
BG 01-02	PoseGait [20]	35.3	47.2	52.4	46.9	45.5	43.9	46.1	48.1	49.4	43.6	31.1	44.5
	GaitGraph [9]	75.8	76.7	75.9	76.1	71.4	73.9	78.0	74.7	75.4	75.4	69.2	74.8
	GaitGraph2 [31]	69.9	75.9	78.1	79.3	71.4	71.7	74.3	76.2	73.2	73.4	61.7	73.2
	GaitSet [5]	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	92.2	94.4	79.0	87.2
	Multimodal [16]	91.9	94.6	96.4	94.3	94.4	91.6	94.1	95.4	95.5	93.9	89.5	93.8
	BiFusion [30]	95.8	97.9	98.2	97.6	94.4	91.6	93.9	96.6	98.5	98.3	93.1	96.0
	Att-CFEN	**85.2**	**86.6**	**89.8**	**86.6**	**87.5**	**85.8**	**85.3**	**89.1**	**89.9**	**90.2**	**83.5**	**87.2**
CL 01-02	PoseGait [20]	24.3	29.7	41.3	38.8	38.2	38.5	41.6	44.9	42.2	33.4	22.5	36.0
	GaitGraph [9]	69.6	66.1	68.8	67.2	64.5	62.0	69.5	65.6	65.7	66.1	64.3	66.3
	GaitGraph2 [31]	57.1	61.1	68.9	66.0	67.8	65.4	68.1	67.2	63.7	63.6	50.4	63.6
	GaitSet [5]	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4
	Multimodal [16]	87.4	96.0	97.0	94.6	94.0	90.1	91.5	94.1	93.8	92.6	88.5	92.7
	BiFusion [30]	88.7	93.9	95.6	93.8	91.4	89.4	92.3	93.8	94.2	93.7	86.2	92.1
	Att-CFEN	**78.6**	**81.4**	**83.5**	**83.1**	**84.0**	**83.2**	**83.1**	**84.4**	**84.0**	**82.7**	**79.0**	**82.5**

4.2.2 Ablation study

To assess the effectiveness of the dual branches of the fusion network in multimodal feature extraction and to evaluate the impact of the constraint attention mechanism in enhancing key gait features, an ablation study was conducted. This experiment involved a sequential analysis: initially

examining the performance of each network branch independently and then evaluating the network's performance with the addition of attention modules, as detailed in Table 2. The specific configurations of the experiment were as follows:

Experiment ① involved the LCFEN subnetwork operating without the LCAM.

Experiment ② exclusively utilized only the JCFEN criteria.

Experiment ③ employed the JCFEN subnetwork, which sans the JCAM.

Experiment ④ operated solely with the LCFEN subnetwork.

Experiment ⑤ involved the Att-CFEN network, omitting the FAM.

Table 2. Ablation study on the CASIA-B dataset

Experimental Modules	Recognition Rate（%）			
	NM	BG	CL	Mean
Experiment ①	89.2	84.7	68.6	80.8
Experiment ②	91.5	85.2	70.1	82.3
Experiment ③	87.3	73.8	67.3	76.1
Experiment ④	89.0	74.5	68.7	77.4
Experiment ⑤	91.8	85.3	80.8	86.0
Att-CFEN	93.0	87.2	82.5	87.6

These configurations were methodically designed to isolate and understand the contributions of each component within the network, thereby elucidating their individual and collective impacts on overall gait feature recognition performance.

The results from the ablation study on the CASIA-B dataset, presented in Table 2, offer insightful findings. The recognition rates for Experiments ② (JCFEN only), ④ (LCFEN only), and the complete Att-CFEN network are 82.3%, 77.4%, and 87.6%, respectively. These outcomes clearly indicate that the Att-CFEN delivers superior overall performance. By integrating both joint motion dependencies and human form features, Att-CFEN uses multimodal and multisemantic data to enrich the representation of high-level gait features. This integration results in a notable increase in the recognition rate by 5.3% and 10.2%, respectively, compared to using the individual feature extraction branches in isolation, underscoring the value of multimodal fusion in enhancing gait feature representation. A comparison among Experiments ①-④ reveals the added value of attention mechanisms. In these setups, the attention mechanisms automatically assign weights to the feature maps, concentrating on distinctive movements and amplifying the representation of significant features. This enhancement leads to improved recognition rates of 1.5% and 1.3%, respectively.

Furthermore, contrasting Experiment ⑤ with the full Att-CFEN network highlights the efficacy of the FAM. This module's role in allocating weights to high-level features from the two modalities results in a 1.6% increase in the recognition rate compared to Experiment ⑤. This finding underscores the importance of differentiating modal data, as they convey unique information. Appropriately weighting these features during fusion enhances the overall recognition rate, emphasizing the significance of nuanced feature integration in multimodal gait analysis.

4.2.3 Multi-Semantic data performance evaluation

For the JCFEN subnetwork, the input data encompass joint coordinates, walking speed, and body sway metrics. To assess how effectively multisemantic data can represent gait features, experiments utilizing various combinations of input data were conducted. The specifics of these combinations, as outlined in Figure 6, are as follows: Combination 1 incorporates silhouette post graphs alongside 3D joint coordinate data. Combination 2 integrated silhouette posture graphs, 3D joint coordinates, joint displacement, and body sway data. These combinations were strategically chosen to explore the incremental value each type of data brings to gait feature expression, thereby shedding light on the utility of multisemantic data in enhancing the accuracy and robustness of gait analysis.

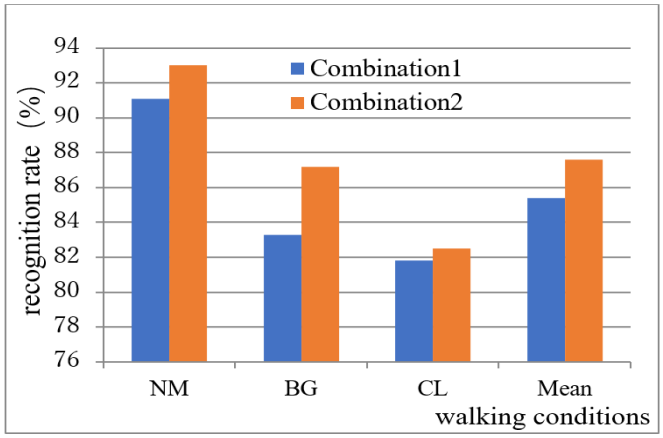

Figure 6. Recognition rate of multisemantic data

In Figure 6, regardless of the walking condition, Combination 2 had the highest recognition rate, indicating that multisemantic data comprehensively express pedestrian gait characteristics from different perspectives, enriching the underlying data features. This allows the Att-CFEN network to focus on different gait characteristics, achieving the best performance.

5. CONCLUSION

The intricate coordination of human joints, bones, and muscles forms the foundation of a pedestrian's gait, characterized by interdependent movements among joints and limbs that create an integrated motion pattern. This paper introduces the Att-CFEN model, a gait-constrained feature mining model equipped with a fusion attention mechanism, focusing on the comprehensive motion constraints between joints. The model begins by addressing the characteristics of the silhouette and skeleton data. A silhouette posture graph is formulated, which, in conjunction with the human skeleton graph, delineates pedestrian gait from two distinct perspectives: motion dependency and handcraft features. To capture the interactive constraints between limbs, local combination graphs depicting limb movement are constructed. The LCFEN subnetwork then abstractly models these limb movements. Similarly, for joint interactions, a constraint adjacency matrix is developed to aggregate the features of joints and their interacting counterparts, with the JCFEN subnetwork modeling these constraint features. To refine the feature extraction process, the LCAM, JCAM, and FAM attention modules are integrated into the two subnetworks and

the multimodal feature fusion stage. These modules play a pivotal role in identifying critical high-level semantic features, calculating the weights of feature maps, and ultimately extracting advanced gait features for pedestrian gait recognition from video data. The efficacy of the proposed Att-CFEN model was rigorously tested using the publicly available CASIA-B dataset. The experimental results demonstrated that the Att-CFEN model notably enhances gait recognition rates, confirming its effectiveness in complex gait analysis and recognition tasks.

FUNDING

This research was funded by Natural Science Basic Research Program of Shaanxi (Grant No.: 2022JM-403).

REFERENCES

[1] Echterhoff, J.M., Haladjian, J., Brügge, B. (2018). Gait and jump classification in modern equestrian sports. In Proceedings of the 2018 ACM International Symposium on Wearable Computers, New York, USA, pp. 88-91. https://doi.org/10.1145/3267242.3267267

[2] Zhang, H., Guo, Y., Zanotto, D. (2019). Accurate ambulatory gait analysis in walking and running using machine learning models. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 28(1): 191-202. https://doi.org/10.1109/TNSRE.2019.2958679

[3] Wang, K., Ding, X., Xing, X., Liu, M. (2019). A survey of multiview gait recognition. Acta Automatica Sinica, 45(5): 841-852.

[4] Sepas-Moghaddam, A., Etemad, A. (2022). Deep gait recognition: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(1): 264-284. https://doi.org/10.1109/TPAMI.2022.3151865

[5] Chao, H., Wang, K., He, Y., Zhang, J., Feng, J. (2021). GaitSet: Cross-view gait recognition through utilizing gait as a deep set. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(7): 3467-3478. https://doi.org/10.1109/TPAMI.2021.3057879

[6] Anusha R, Jaidhar C.D. (2020). Clothing invariant human gait recognition using modified local optimal oriented pattern binary descriptor. Multimedia Tools and Applications, 79(3): 2873-2896.

[7] Li, S., Sun, P., Lang, Y.B. (2020). Research on gait recognition method in surveillance video based on LSTM. Journal of People's Public Security University of China (Science and Technology), 26(2): 23-28.

[8] Lin, B., Zhang, S., Yu, X. (2021). Gait recognition via effective global-local feature representation and local temporal aggregation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, Canada, pp. 14648-14656.

[9] Teepe, T., Khan, A., Gilg, J., Herzog, F., Hörmann, S., Rigoll, G. (2021). Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, pp. 2314-2318. https://doi.org/10.1109/ICIP42928.2021.9506717

[10] Song, Y.F., Zhang, Z., Shan, C., Wang, L. (2020). Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. In Proceedings of the 28th ACM International Conference on Multimedia, Dalian, China, pp. 1625-1633. https://doi.org/10.1145/3394171.3413802

[11] Tian, H.Y., Ma, X., Li, Y.B. (2022). Skeleton-based abnormal gait recognition: A survey. Journal of Jilin University (Engineering and Technology Edition), 52(4): 725-737.

[12] Fan, C., Peng, Y., Cao, C., Liu, X., Hou, S., Chi, J., Huang, Y., Li, Q., He, Z. (2020). Gaitpart: Temporal part-based model for gait recognition. In Proceedings of The IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14225-14233.

[13] Uddin, M.Z., Muramatsu, D., Takemura, N., Ahad, M.A.R., Yagi, Y. (2019). Spatio-temporal silhouette sequence reconstruction for gait recognition against occlusion. IPSJ Transactions on Computer Vision and Applications, 11(1): 1-18. https://doi.org/10.1186/s41074-019-0061-3

[14] Seong, S., Choi, J. (2021). Semantic segmentation of urban buildings using a high-resolution network (HRNet) with channel and spatial attention gates. Remote Sensing, 13(16): 3087. https://doi.org/10.3390/rs13163087

[15] Cao, Z., Simon, T., Wei, S.E., Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, USA, pp. 7291-7299.

[16] Wang, L., Chen, J. (2022). A two-branch neural network for gait recognition. arXiv Preprint arXiv: 2202.10645, 3(6): 7.

[17] Wu, Z., Huang, Y., Wang, L. (2015). Learning representative deep features for image set analysis. IEEE Transactions on Multimedia, 17(11): 1960-1968. https://doi.org/10.1109/TMM.2015.2477681

[18] Li, C., Min, X., Sun, S., Lin, W., Tang, Z. (2017). DeepGait: A learning deep convolutional representation for view-invariant gait recognition using joint Bayesian. Applied Sciences, 7(3): 210. https://doi.org/10.3390/app7030210

[19] Chai, T., Mei, X., Li, A., Wang, Y. (2021). Silhouette-based view-embeddings for gait recognition under multiple views. In 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, pp. 2319-2323. https://doi.org/10.1109/ICIP42928.2021.9506238

[20] Liao, R., Yu, S., An, W., Huang, Y. (2020). A model-based gait recognition method with body pose and human prior knowledge. Pattern Recognition, 98: 107069. https://doi.org/10.1016/j.patcog.2019.107069

[21] Qi, Y.J., Kong, Y.P., Wang, J.J., Zhu, X.D. (2021). Gait recognition method combining LSTM and CNN. Journal of Xidian University, 48(05): 78-85.

[22] Thapar, D., Nigam, A., Aggarwal, D., Agarwal, P. (2018). VGR-net: A view invariant gait recognition network. In 2018 IEEE 4th International Conference on Identity, Security, and Behavior Analysis (ISBA), Singapore, pp. 1-8. https://doi.org/10.1109/ISBA.2018.8311475

[23] Li, S., Sun, P., Lang, Y.B. (2020). Research on gait recognition method in surveillance video based on LSTM. Journal of People's Public Security University of China (Science and Technology), 26(2): 23-28.

[24] Wolf, T., Babaee, M., Rigoll, G. (2016). Multi-view gait recognition using 3D convolutional neural networks. In

2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, pp. 4165-4169. IEEE. https://doi.org/10.1109/ICIP.2016.7533144

[25] Hou, S., Cao, C., Liu, X., Huang, Y. (2020), Gait lateral network: Learning discriminative and compact representations for gait recognition. In Proceedings of ECCV, Glasgow, UK, 2020, pp.382-398.

[26] Yan, S., Xiong, Y., Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of The AAAI Conference on Artificial Intelligence, 32(1). https://doi.org/10.1609/aaai.v32i1.12328

[27] Hasan, M.B., Ahmed, T., Kabir, M.H. (2022). Heatgait: Hop-extracted adjacency technique in graph convolution based gait recognition. In 2022 4th International Conference on Advances in Computer Technology, Information Science and Communications (CTISC), IEEE, Suzhou, China, pp. 1-6. https://doi.org/10.1109/CTISC54888.2022.9849799

[28] Castro, F.M., Marin-Jimenez, M.J., Guil, N., Pérez de la Blanca, N. (2020). Multimodal feature fusion for CNN-based gait recognition: an empirical comparison. Neural Computing and Applications, 32: 14173-14193. https://doi.org/10.1007/s00521-020-04811-z

[29] Li, G., Guo, L., Zhang, R., Qian, J., Gao, S. (2023). TransGait: Multimodal-based gait recognition with set transformer. Applied Intelligence, 53(2): 1535-1547. https://doi.org/10.1007/s10489-022-03543-y

[30] Peng, Y., Ma, K., Zhang, Y., He, Z. (2024). Learning rich features for gait recognition by integrating skeletons and silhouettes. Multimedia Tools and Applications, 83(3): 7273-7294. https://doi.org/10.1007/s11042-023-15483-x

[31] Teepe, T., Gilg, J., Herzog, F., Hörmann, S., Rigoll, G. (2022). Towards a deeper understanding of skeleton-based gait recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Louisiana, USA, pp. 1569-1577.