Journal of Clinical and Nursing Research, 2025, Volume 9, Issue 11
http://ojs.bbwpublisher.com/index.php/JCNR
Online ISSN: 2208-3693
Print ISSN: 2208-3685

Skeleton–Silhouette Complementary Perception: Toward Robust Gait Recognition

Xiaokai Liu*, Luyuan Hao

Information Science and Technology College, Dalian Maritime University, Dalian 116026, Liaoning Province, China

**Corresponding author: Xiaokai Liu, xkliu@dlmu.edu.cn*

Abstract: Gait, the unique pattern of how a person walks, has emerged as one of the most promising biometric features in modern intelligent sensing. Unlike fingerprints or facial characteristics, gait can be captured unobtrusively and at a distance, without requiring the subject's awareness or cooperation. This makes it highly suitable for long-range surveillance, forensic investigation, and smart environments where contactless recognition is crucial. Traditional gait-recognition systems rely either on silhouettes, which capture the outer appearance of a person, or on skeletons, which describe the internal structure of human motion. Each modality provides only a partial understanding of gait. Silhouettes emphasize shape and contour but are easily distorted by clothing or carried objects; skeletons describe motion dynamics and limb coordination but lose discriminative details about body shape. This article presents the concept of Complementary Semantic Embedding (CSE), a unified framework that merges silhouette and skeleton information into a comprehensive semantic representation of human walking. By modeling the complementary nature of appearance and structure, the approach achieves more robust and accurate gait recognition even under challenging conditions.

Keywords: Complementary perception; Gait recognition; Feature fusion

Online publication: Dec 12, 2025

1. Introduction: Why Gait Matters

Among all biometric traits, gait occupies a special position [1]. It can be observed at a distance, through ordinary video cameras, without any physical contact. Every individual walks with a distinctive rhythm and spatial pattern, shaped by anatomy, age, and lifestyle. For security systems, this means people can be recognized even when their faces are hidden or when they move through crowded environments. For healthcare applications, changes in gait can reveal early signs of neurological or muscular disorders.

Yet, identifying a person by the way they walk is far from trivial [2]. Human walking involves a highly complex coordination of limbs and joints, influenced by countless factors such as clothing, footwear, walking speed, load carrying, and camera perspective. Two video sequences of the same person may look entirely

different simply because of lighting or attire.

Early computer-vision systems focused primarily on silhouette analysis: extracting the outline of a walking person from each video frame, aligning the silhouettes over time, and using them as dynamic templates for classification. These silhouettes contain detailed information about body shape, but can change drastically when a person wears a long coat or carries a backpack. Later, with the progress of human pose estimation technology, another representation emerged-the skeleton graph, composed of discrete key joints such as shoulders, hips, knees, and ankles. This abstraction captures the geometry and motion of the human body, largely independent of appearance. It provides robustness against clothing variation but discards subtle cues about physique that help distinguish individuals. Each of these representations tells part of the story; neither alone can fully describe the individuality of gait. The key question therefore, becomes: how can we combine the strengths of both silhouettes and skeletons to form a complete picture of human walking?

2. From silhouettes to skeletons: Principles of gait recognition

To understand how gait can serve as a biometric, it helps to visualize the process. Imagine a sequence of video frames showing a person walking across a corridor. A typical gait recognition system performs three major steps:

(1) Detection and Preprocessing: The person is isolated from the background. Silhouette extraction methods produce binary masks outlining the figure; pose-estimation algorithms detect skeletal joints.

(2) Feature Representation: These dynamic signals are transformed into numerical features. For silhouettes, this might be a Gait Energy Image, an averaged image summarizing a walking cycle. For skeletons, features include joint coordinates and limb angles over time.

(3) Classification and Recognition: Machine-learning models analyze these features to identify the person or verify identity.

Silhouette-based methods capture appearance dynamics, providing information about limb shape and proportion. They are powerful when the subject's clothing and lighting conditions remain consistent. However, they fail when visual cues change. Skeleton-based methods capture structural dynamics, tracking relative motion between joints. These are invariant to external appearance but sometimes ambiguous - two individuals of different builds may share very similar skeleton patterns. The challenge, then, is how to merge these two complementary views so that the system understands both what a person looks like and how they move.

3. The idea of complementary semantic embedding

The Complementary Semantic Embedding (CSE) framework begins with a simple but profound observation: silhouette and skeleton data describe the same physical event - a person walking - but from two different perspectives. Like two languages describing the same story, they each express unique semantics. Integrating them can reveal meaning that is hidden when either is viewed in isolation. The CSE model treats silhouettes and skeletons as homologous isomerism data - two forms of signals derived from the same source. It then aligns and fuses them at a high semantic level rather than at the raw data or feature level. This semantic fusion allows the system to learn correlations between external shape and internal structure.

In essence, the method aims to capture three key complementarities:

(1) Robustness compensation: When silhouettes are distorted by clothing or accessories, skeletons provide invariant structural cues.

(2) Appearance compensation: When skeletons lack discriminative body-shape details, silhouettes restore identity-related information.

(3) Dynamic compensation: When self-occlusion hides part of a silhouette, skeleton motion data recover missing dynamics. By combining these aspects, the framework achieves a more balanced understanding of gait - one that is both structurally consistent and appearance-aware.

4. How the framework works

4.1. Concept overview

The Complementary Semantic Embedding system consists of three major components:

(1) Data preprocessing: Converts raw videos into synchronized silhouette and skeleton streams.

(2) Dual feature extraction: Uses specialized neural networks to process each modality.

(3) Complementary fusion and classification: Merges the two streams into a unified representation for final recognition.

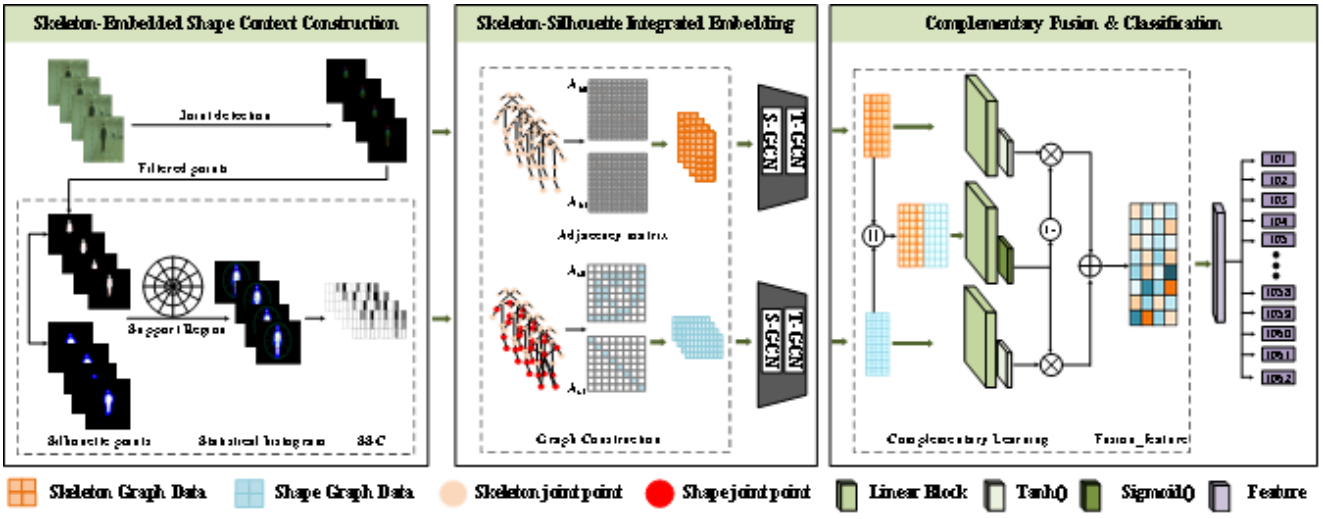

Figure 1. The Complementary Semantics Embedded Framework includes three modules: Data preprocessing, Specific convolution network, and Complementary feature fusion for classification.

4.2 Skeleton-Embedded Shape Context (SSC)

At the heart of the method lies the Skeleton-Embedded Shape Context (SSC), a representation that connects key skeletal joints with their surrounding silhouette regions. Imagine tracing the outline of a person's silhouette and marking the location of key joints like knees or elbows. Around each joint, a set of local regions is defined to describe how the body shape extends outward. These regions act like a "field" linking internal bones with external appearance. This hybrid representation allows the model to learn, for example, how the curvature of a leg silhouette relates to the bending angle of the knee joint. Through such correspondences, the network develops a deeper physical understanding of motion, capturing what moves and how it moves simultaneously.

4.3. Complementary Semantic Embedding

Once both streams, silhouette features and skeleton features, are extracted, the system employs an attention-based fusion mechanism. This module evaluates, dimension by dimension, which modality contributes more discriminative information and assigns adaptive weights accordingly. In regions where silhouette information is reliable (for instance, the outline of a leg), it dominates; where skeleton data are more stable (like under occlusion), the system leans on structural cues.

Mathematically, this corresponds to a weighted combination of nonlinear transformations, guided by learned importance scores. Conceptually, it resembles a conversation between two experts: one describing outer form, the other internal mechanics. The attention unit ensures their opinions combine intelligently into a single judgment.

4.4. Intuitive flow of the framework

Video frames enter two parallel paths, silhouette analysis and skeleton analysis, each producing a stream of high-level features. These converge within a fusion module that learns to integrate complementary semantics. The resulting vector represents a person's gait identity in a compact, discriminative form.

5. Experimental validation and key findings

To evaluate this approach, researchers tested it on a major benchmark datasets: CASIA-B - collected by the Chinese Academy of Sciences, containing 124 subjects walking under three conditions: normal, with a bag, and with a coat, from 11 camera angles. In This case, skeletons were extracted using OpenPose [3], and silhouettes were preprocessed from video frames. The experiments compared the proposed framework with well-known baseline models relying on single modalities.

5.1. Robustness and accuracy

Across all conditions, the complementary framework significantly outperformed traditional methods, including SPAE GV [4], GaitGANv1 [5], GaitGANV2 [6], MGANs [7], PoseGait [8]. In the normal walking scenario, accuracy improved to above 90 percent; in more challenging carrying-bag or wearing-coat cases, the improvement margin exceeded 20 percent over silhouette-only models. The results confirm that skeleton cues help maintain identity consistency even when appearance changes. Conversely, silhouettes enhance recognition when skeleton extraction is uncertain due to occlusion or camera angle.

Table 1. Comparison of identification accuracy with comparable methods

Gallery NM 1-4						**0–180**						**Mean**
Prob												
SPAE GV [4]	50.0	58.6	61.1	63.3	64.0	62.1	62.3	66.3	64.4	54.5	46.7	62.8
GaitGANv1 [5]	41.9	53.6	63.0	64.5	63.2	58.2	61.7	65.7	62.7	54.1	40.6	61.0
GaitGANV2 [6]	48.1	61.9	68.7	71.7	66.7	64.8	66.1	70.2	71.6	56.8	46.1	66.2
MGANs [7]	54.9	65.9	72.1	74.8	71.1	65.7	70.0	75.6	76.2	68.6	53.8	68.1
PoseGait [8]	49.7	61.6	67.0	66.7	60.8	59.0	62.5	61.4	67.3	62.0	62.0	63.7
Ours	87.2	91.7	94.0	92.6	87.9	87.0	88.4	91.9	91.7	88.4	83.2	90.1

5.2. Feature distribution insights

Using dimensionality-reduction visualization (t-SNE), the fused representations formed compact clusters for each identity, with clear separation between individuals. This pattern indicates that the model learned an effective embedding space where people are easily distinguishable despite covariate variations.

6. Applications and broader implications

The implications of robust multimodal gait recognition extend far beyond identification tasks.

6.1. Security and surveillance

In public security, gait recognition enables continuous monitoring without violating privacy, since it relies on motion rather than facial details. Systems deployed in airports, subways, or public buildings could detect known individuals or abnormal behaviors automatically. The fusion approach described here makes such systems more reliable in real-world conditions where people wear different clothes or carry personal items.

6.2. Healthcare and rehabilitation

Gait is a sensitive indicator of human health. Neurological disorders such as Parkinson's disease, stroke, or muscular dystrophy manifest early in walking irregularities. By combining structural and appearance cues, the complementary perception model can detect subtle changes in movement patterns that might escape the human eye. Integrating such models into wearable or ambient sensors could revolutionize preventive healthcare.

6.3. Beyond vision: Toward multimodal human sensing

The principles behind complementary perception extend naturally to other sensing modalities - radar, infrared, sonar, and wireless radio. Each captures different physical aspects of human motion. Integrating them could lead to general multimodal perception networks capable of understanding human activity in almost any environment, day or night.

7. Outlook: Toward intelligent human sensing

The Complementary Semantic Embedding framework represents a step toward holistic human understanding. Future research will likely focus on three directions:

(1) Cross-modal generalization: Extending the framework to incorporate non-visual data such as millimeter-wave radar or depth sensors. These can penetrate clothing or operate in darkness, complementing visual modalities.

(2) Lightweight deployment: Optimizing the architecture for edge computing and embedded AI chips so that gait recognition runs in real time on mobile or IoT devices.

(3) Explainable perception: Interpreting what the fused features mean physically, which aspects of motion define identity, and how they relate to biomechanics or emotion.

Ultimately, the goal is to build intelligent sensing systems that can perceive and reason about human behavior comprehensively - not merely as pixels or coordinates, but as expressions of identity, health, and intention. The fusion of silhouette and skeleton perception provides a tangible example of how AI can integrate

heterogeneous information to achieve a deeper level of understanding.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Han J, Bhanu B, 2006, Individual Recognition Using Gait Energy Image. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(2): 316–322.

[2] Wu Z, Huang Y, Wang L, ET AL., 2017, A Comprehensive Study on Cross-View Gait Based Human Identification With Deep CNNs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(2): 209–226.

[3] Cao Z, Hidalgo G, Simon T, et al., 2021, OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(1): 172–186.

[4] Yu S, Chen H, Wang Q, et al., 2017, Invariant Feature Extraction for Gait Recognition Using Only One Uniform Model. Neurocomputing, 239: 81–93.

[5] Yu S, Chen H, Reyes E, et al., 2017, GaitGAN: Invariant Gait Feature Extraction Using Generative Adversarial Networks. IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017: 532–539.

[6] Yu S, Liao R, An W, et al., 2019, GaitGANv2: Invariant Gait Feature Extraction Using Generative Adversarial Networks. Pattern Recognition, 2019: 179–189.

[7] He Y, Zhang J, Shan H, et al., 2019, Multi-Task GANs for View-Specific Feature Learning in Gait Recognition. IEEE Transactions on Information Forensics and Security, 2019: 102–113.

[8] Liao R, Yu S, An W, et al., 2019, A Model-Based Gait Recognition Method With Body Pose and Human Prior Knowledge. Pattern Recognition, 98: 107–169.