International Journal of Advanced Technology in Engineering and Science
Vol. No. 14, Issue No. 04, April 2026
www.ijates.com
ijates
ISSN 2348 - 7550

Covariate Robust Gait Recognition Using CNN and BiLSTM

K. V. Satyanarayana[1], V Vijaya Santhi[2]

[1,2]Department of Electronics and Communication Engineering

JNTU-GV College of Engineering, Vizianagaram, Andhra Pradesh, India

Abstract

Human gait recognition is a behavioral biometric technique capable of identifying individuals at a distance without requiring subject cooperation. Unlike fingerprint or iris recognition, it can function using ordinary video surveillance systems, making it valuable for security and forensic applications. However, gait recognition performance degrades sharply under covariate conditions such as clothing changes, carried objects, and viewpoint variation. This paper proposes a covariate-robust, multimodal gait recognition framework that fuses silhouette and pose information through a novel Skeleton-Enhanced Silhouette (S-Silhouette) representation. A dual-stream 2D Convolutional Neural Network (CNN) extracts complementary spatial features from the silhouette and S-Silhouette streams, which are fused and passed to a Bidirectional Long Short-Term Memory (BiLSTM) network for temporal modeling. The network is trained with a Triplet Loss objective to learn discriminative gait embeddings. Experiments on the CASIA-B benchmark dataset show that the proposed two-stream CNN+BiLSTM model attains 100% accuracy under normal walking, 83% under bag-carrying, and 91% under clothing-variation conditions, outperforming silhouette-only, pose-only, and simple fusion baselines, particularly under covariate variation.

Keywords: *Gait Recognition, Biometrics, CNN, BiLSTM, Triplet Loss, S-Silhouette, CASIA-B, Covariate Robustness*

I. Introduction

Biometrics refers to the automated recognition of individuals based on distinctive physiological and behavioral characteristics. Compared with physiological traits such as fingerprints or iris patterns, gait is a behavioral biometric derived from an individual's walking pattern, influenced by body structure and habitual movement. A key advantage of gait recognition is that it can operate at a distance, without subject cooperation, and using standard video surveillance infrastructure, making it well suited to public surveillance, forensic analysis, and access control.

Despite these advantages, gait recognition systems remain sensitive to covariate conditions — clothing changes, carried objects, walking speed, and camera viewpoint — which alter the visual appearance and motion pattern of a subject and significantly degrade silhouette-based representations. Most existing methods rely on a single modality (either silhouette or pose), which limits their ability to capture the full characteristics of human gait and to remain robust under real-world variation.

This work proposes a covariate-resilient, multimodal gait recognition framework that combines silhouette and pose information through a unified S-Silhouette representation, and models spatial and

temporal characteristics jointly using a dual-stream CNN and a BiLSTM network, trained with a Triplet Loss metric-learning objective. The main objectives of this work are to: (i) detect and identify individuals under varying clothing and carrying conditions, (ii) use deep learning to extract discriminative gait features, and (iii) fuse complementary gait features to improve recognition accuracy and robustness.

II. Related Work

Early gait recognition research relied on handcrafted, silhouette-based representations. Han and Bhanu introduced the Gait Energy Image (GEI), formed by averaging silhouettes over a gait cycle, which compactly encodes spatio-temporal information but remains sensitive to clothing and carrying variations. Liu and Sarkar used probabilistic Hidden Markov Models to normalize gait dynamics under variation, while Guan et al. proposed classifier ensembles for covariate-resistant recognition across view and clothing changes.

Pose-based approaches model the internal skeletal structure of the body and are comparatively more robust to appearance change, though dependent on accurate joint detection. Kathing et al. combined GEI with gradient and local binary pattern features for robustness against clothing and carrying variation, and Zou et al. fused RGB-D and inertial sensor data for real-world robustness. More recent deep learning approaches such as GaitSet, GaitPart, MetaGait, and hierarchical gait models leverage automatic feature learning and part-based or meta-learning strategies to improve accuracy, but most still rely on a single feature modality, leaving a gap for approaches that jointly exploit silhouette shape and skeletal dynamics along with temporal modeling.

III. Proposed Methodology

A. S-Silhouette Representation

To overcome the limitations of standalone silhouette and pose representations, this work introduces the S-Silhouette (Skeleton-Enhanced Silhouette), which embeds skeletal joint data within silhouette frames. Joint coordinates obtained from a pre-trained pose estimation model are projected onto the same spatial dimensions as the silhouette mask, and skeletal connections are drawn to form a structured body representation. This hybrid representation preserves the external body contour while encoding internal joint positions and movement, improving robustness to clothing and appearance changes.

B. System Architecture

Raw RGB video frames are processed into two parallel streams: (1) binary silhouette sequences obtained via background subtraction, morphological refinement, and normalization to a fixed resolution, and (2) S-Silhouette sequences generated by projecting validated pose keypoints onto the silhouette mask. A dual-stream 2D CNN independently extracts hierarchical spatial features from each

modality — one stream capturing global body shape and stride structure, the other capturing skeletal joint dynamics.

The two feature streams are aligned and combined using feature-level fusion (concatenation followed by fully connected layers), producing a unified spatio-structural descriptor per frame. The fused temporal sequence is passed to a Bidirectional LSTM, which processes the sequence in forward and backward directions to capture stride rhythm, periodicity, and inter-frame dependencies. Temporal Average and Max Pooling are applied to the BiLSTM outputs to produce a fixed-length gait embedding (256–512 dimensions) that serves as the subject's biometric signature.

C. Triplet Loss Training

The network is trained using a Triplet Loss objective operating on (Anchor, Positive, Negative) samples, defined as:

$$L = max(0, d(a, p) - d(a, n) + \alpha)$$

where d(·) is the Euclidean distance between embeddings and α is a margin hyperparameter. The objective minimizes the distance between embeddings of the same subject (anchor–positive) while maximizing the distance to a different subject (anchor–negative). Hard-negative mining (Batch-Hard sampling) is used to select informative triplets, forcing the network to learn kinematic features that remain invariant to appearance changes rather than relying on superficial similarity.

D. Multi-View Learning

To achieve robustness to camera viewpoint, training batches are constructed so that a single identity is represented by sequences from multiple, disparate viewing angles (0°–180° in the CASIA-B dataset). Because the Triplet Loss penalizes failure to cluster same-identity embeddings regardless of viewing angle, the network is compelled to learn view-invariant kinematic correlations rather than view-specific spatial artifacts, and is evaluated using a cross-view gallery/probe protocol.

IV. Experimental Results and Discussion

A. Dataset and Setup

The proposed framework is implemented in Python using PyTorch, with OpenCV for video/silhouette preprocessing and TensorFlow Hub MoveNet for pose estimation. Evaluation is performed on the CASIA-B benchmark dataset, which provides gait sequences under three walking conditions — Normal (NM), Bag-carrying (BG), and Clothing variation (CL) — across 11 viewing angles, enabling assessment of both covariate and viewpoint robustness.

B. Training Behavior

Training loss decreases steadily over 200 epochs, from an initial value near 0.8 to below 0.1, indicating stable convergence of the Triplet Loss objective and the absence of divergence or sharp oscillation, confirming that the selected learning rate, optimizer (Adam), and batch size are well suited to the task.

C. Comparative Performance

Table I compares the proposed two-stream CNN+BiLSTM model against silhouette-only (GEI), pose-only (PEI), and simple fusion baselines under the three CASIA-B walking conditions.

Model	NM (%)	BG (%)	CL (%)
GEI (2D CNN)	100	75	78
PEI (2D CNN)	60	35	55
GEI + PEI Fusion (2D CNN)	100	65	45
S-Silhouette Energy Image (2D CNN)	100	55	45
Condition-Invariant Triplet (CNN + BiLSTM)	100	78	75
Proposed Two-Stream Silhouette + Pose (2×CNN + BiLSTM)	100	83	91

TABLE I. RECOGNITION ACCURACY (%) UNDER NORMAL (NM), BAG-CARRYING (BG), AND CLOTHING (CL) CONDITIONS

All models achieve 100% accuracy under normal walking, where gait patterns are unobstructed. Under bag-carrying, the proposed two-stream model attains 83% accuracy, exceeding the condition-invariant triplet baseline (78%) and the GEI-only model (75%). The largest gain is observed under clothing variation, where the proposed model reaches 91% accuracy — a substantial improvement over the GEI (78%), PEI (55%), GEI+PEI fusion (45%), and S-Silhouette-only (45%) baselines, confirming the benefit of jointly modeling spatial fusion and temporal dynamics rather than relying on a single-frame fused representation alone.

D. Ablation Study

Removing the S-Silhouette stream reduces accuracy, confirming its role in capturing combined spatial and structural cues. Excluding feature fusion prevents the model from exploiting complementary silhouette and pose information, and removing the BiLSTM temporal modeling stage reduces the model's ability to capture sequential dependencies such as stride rhythm — all three components are shown to contribute significantly to overall performance.

E. Discussion

The results confirm that combining silhouette and pose information through the S-Silhouette representation, together with BiLSTM-based temporal modeling and Triplet Loss embedding learning, substantially improves robustness to covariate conditions relative to single-modality or purely spatial baselines. Performance under extreme occlusion or highly irregular gait, and the computational cost of the dual-stream architecture, remain limitations for real-time deployment.

V. Conclusion and Future Work

This paper presented a covariate-robust, multimodal gait recognition framework built around a novel S-Silhouette representation, a dual-stream CNN for spatial feature extraction, feature-level fusion, BiLSTM-based temporal modeling, and Triplet Loss embedding learning. Evaluated on the CASIA-B

dataset, the proposed model achieved 100% accuracy under normal walking and improved robustness of 83% and 91% under bag-carrying and clothing-variation conditions respectively, outperforming silhouette-only, pose-only, and naïve fusion baselines. Future work will explore transformer-based attention mechanisms, real-time optimization for surveillance deployment, cross-dataset generalization, and extension to 3D gait representations to further improve robustness under extreme covariate and viewpoint variation.

References

[1] C. Hong, J. Yu, D. Tao, and M. Wang, "Image-based three-dimensional human pose recovery by multiview locality-sensitive sparse retrieval," IEEE Trans. Ind. Electron., vol. 62, no. 6, pp. 3742–3751, 2014.

[2] C. Hong, J. Yu, J. Wan, D. Tao, and M. Wang, "Multimodal deep autoencoder for human pose recovery," IEEE Trans. Image Process., vol. 24, no. 12, pp. 5659–5670, 2015.

[3] D. Kastaniotis, I. Theodorakopoulos, and S. Fotopoulos, "Pose-based gait recognition with local gradient descriptors and hierarchically aggregated residuals," J. Electron. Imaging, vol. 25, no. 6, 063019, 2016.

[4] R. Liao, S. Yu, W. An, and Y. Huang, "A model-based gait recognition method with human prior knowledge," Pattern Recognit., vol. 98, 107069, 2020.

[5] E. R. Isaac et al., "Template-based gait authentication through Bayesian thresholding," IEEE/CAA J. Autom. Sinica, vol. 6, no. 1, pp. 209–219, 2019.

[6] Y. Zhang et al., "Cross-view gait recognition by discriminative feature learning," IEEE Trans. Image Process., vol. 29, pp. 1001–1015, 2019.

[7] X. Chen et al., "Multi-view gait image generation for cross-view gait recognition," IEEE Trans. Image Process., 2020.

[8] S. Han and B. Bhanu, "Individual recognition using gait energy image," IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 2, pp. 316–322, 2006.