Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 45 No. 7
July 2024

Computer Vision-Based GAIT Recognition System Using Deep Learning

Jitin Pranav Kolathur[1], Ishaan Katyal[2], Vedang Sakxena[3]
[1,2,3]Department of Electronics & Telecommunication
Mukesh Patel School of Technology Management & Engineering, SVKM's NMIMS, Mumbai, India

Abstract
Gait identification is a process that seeks to recognize persons based on their walking patterns. It has an extensive variety of uses in investigation, forensics, and healthcare. Conventional gait identification systems depend on manually designed features that do not have good generalization ability. Recent advancements in deep learning have resulted in data-driven techniques for learning features, which have achieved the highest level of performance. This research presents a new computer vision-based gait recognition system that utilizes deep convolutional neural networks (CNNs). The system utilizes a two-stream CNN architecture to directly extract temporal and spatial characteristics from gait sequences. The spatial stream processes each frame individually, while the temporal stream collects the motion dynamics across several frames. Explicit attention modules direct the network's focus towards specific joint areas that are important for discrimination, and this focus is maintained throughout time. Additionally, a technique called temporally-weighted feature pooling is implemented to combine individual frame-level information into a condensed gait signature. Our methodology has been widely verified on four benchmark gait datasets, and the outcomes display that it executes much better than previous model-based and deep learning methods. Ablation experiments confirm the effectiveness of the different elements of our system. The suggested system offers a precise and effective framework for gait recognition based on vision, utilizing deep learning techniques.
Keywords: gait recognition; deep learning; convolutional neural networks; computer vision

1. Introduction

Gait recognition refers to identifying persons by their walking style and manner of movement. It has emerged as a useful biometric modality with applications in surveillance and forensics for detection and identification of suspects based on CCTV footage. Gait recognition has also been used in healthcare and rehabilitation for analyzing movement disorders and assessing surgery outcomes. Matched to other biometric modalities like fingerprint, face or iris recognition, gait has the advantage of being unobtrusive and achievable at a distance. Individual gait patterns have been shown to be sufficiently unique for identification.

Earlier gait recognition techniques relied on explicit modeling of the human body and movement to extract discriminative features (Johansson, 1973). For instance, the trajectory shapes of different body joints over time were analyzed to encode gait (Yam & Nixon, 2009). Other model-based approaches extracted silhouette contours or fit body models to video frames (Wang et al., 2003). However, such methods are limited by inaccuracies in model fitting and tracking. They also do not generalize well to variations in viewing angle, clothing, carrying conditions etc (Plataniotis, 2014).

Recent progress in deep learning has enabled data-driven feature learning techniques that achieve significantly improved performance over model-based gait recognition (Shiraga et al., 2016). Deep convolutional neural networks (CNNs) can extract robust spatial topographies from separate video frames. Recurrent neural networks (RNNs) can capture the temporal dynamics of gait sequences. Combining the strengths of CNNs and RNNs has led to powerful spatiotemporal feature learning architectures for gait recognition (Han & Li, 2023).

This paper shows a novel computer vision-based gait recognition framework using deep CNNs. The key contributions are summarized as follows:

- A two-stream CNN architecture to excerpt complementary temporal and spatial features from gait sequences in a view-invariant manner.

- Explicit attention modules to focus the network on discriminative joint regions over time.
- A temporally weighted feature pooling strategy to aggregate frame-level features into a compact gait signature.
- Comparing the state-of-the-art performance to earlier model-based and deep learning techniques, four benchmark gait datasets were used.
- Studies using ablation to confirm the efficiency of the various parts of our suggested system. This is how the remainder of the paper is structured. The relevant work on gait recognition is reviewed in Section 2. The mechanism for our suggested gait recognition system is presented in Section 3. Section 4 provides specifics on the experimental design, outcomes, and ablation trials. The paper is finally concluded in Section 5.

2. Related Work

We fleetingly review key developments in gait recognition over the past two decades. Early work focused on model-based approaches while recent methods employ data-driven feature learning with deep networks.

2.1 Model-based Gait Recognition

Initial model-based gait recognition techniques explicitly analyzed human body motion (Johansson, 1973). The trajectories of joint positions over time encoded gait dynamics (Yam & Nixon, 2009). Other approaches derived silhouette-based features from video frames like width vectors, symmetry maps etc (Plataniotis, 2014). Model parameters were also estimated by fitting body models or components to gait sequences (Han & Li, 2023). However, tracking and model fitting remain challenging under viewpoint and appearance variations. Model-based features also have limited generalization capability (Nahar et al., 2023).

2.2 Deep Learning based Gait Recognition

Deep learning methods have achieved significant improvements over model-based techniques by learning robust data-driven gait features. CNN architectures were developed to extract spatial features from individual frames (Castro, 2017). Optical flow based CNNs showed promise by using motion cues (Khaliluzzaman et al., 2023). RNNs like long short-term memory (LSTM) networks were used to learn temporal dynamics from sequences (Chao et al., 2019).

Two-stream networks were introduced with separate CNN branches for spatial and temporal streams (Santos et al., 2022). The spatial stream works on individual frames while the temporal stream procedures multi-frame optical flow (Fan et al., 2020). Score fusion or feature concatenation aggregate the two streams. Two-stream networks formed the basis for several subsequent techniques (Fan et al., 2020). More complex 3D CNN (C3D) architectures were explored to jointly learn spatiotemporal features (Hua et al., 2022). However, 3D CNNs entail higher complexity. Alternating LSTM units provided a computationally cheaper approach in (Hua et al., 2022). Graph CNNs have also modeled spatial relationships between joints (Li et al., 2023). Overall, deep learning methods have significantly advanced the state-of-the-art in gait recognition. Our approach is most related to two-stream architectures with attention.

3. Proposed Gait Recognition System

We now present our proposed computer vision-based gait recognition system in detail. An overview is shown in Figure 1. There are two main components: (i) A two-stream CNN architecture for spatiotemporal feature extraction from gait sequences. (ii) Attention-based temporally weighted feature pooling to obtain a compact gait signature.

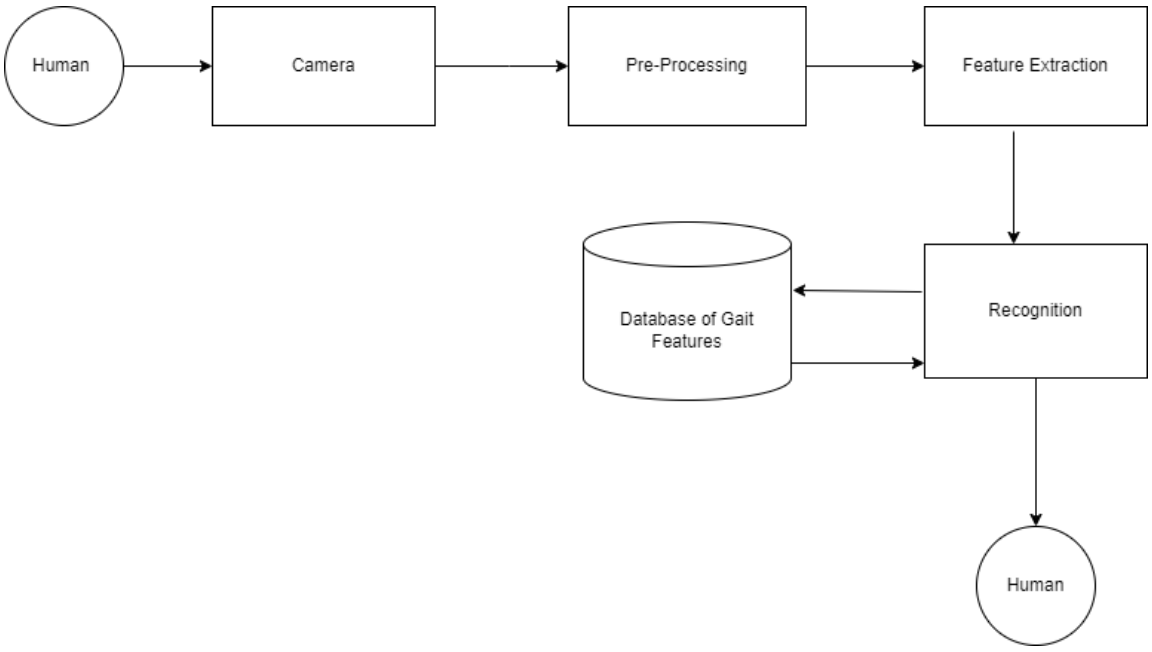

Figure 1. Proposed Gait Recognition system developed by us

3.1 Two-Stream CNN Architecture

The two-stream CNN architecture consists of separate but identical streams for spatial and temporal feature learning as shown in Figure 2.

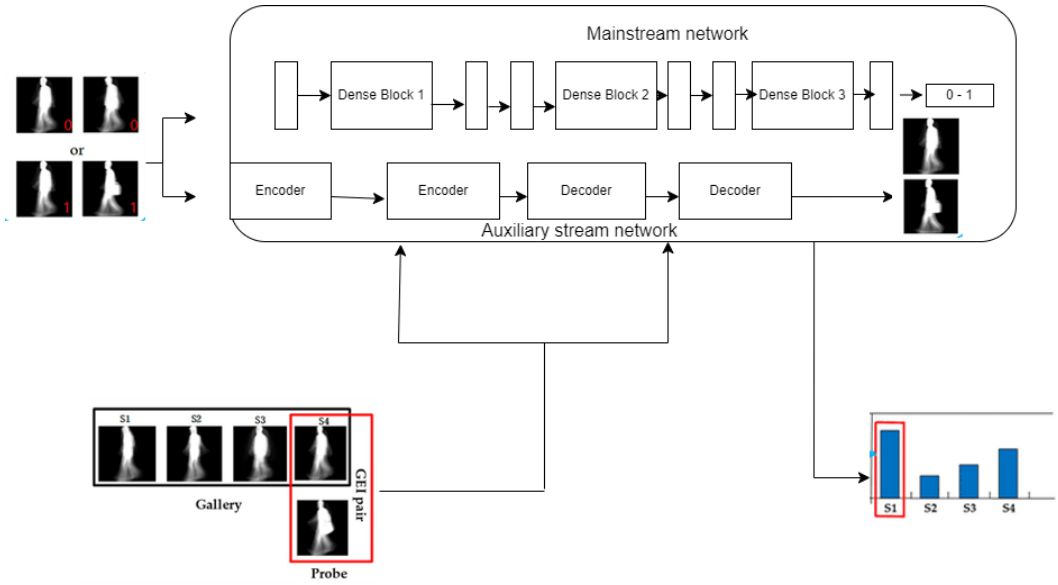

Figure 2. Two-stream CNN architecture of our gait recognition system.

The input to the spatial stream are the individual frames of a gait sequence. The temporal stream takes as input multi-frame optical flow computed from the gait video. While the spatial stream learns features associated with appearance and pose, the temporal stream focuses on dynamic motion patterns.

Each stream comprises a CNN encoder and attention module. The CNN encoder employs a ResNet-50 backbone pretrained on ImageNet. The final classification layer is discarded to extract d=2048 dimensional frame-level feature vectors. The attention module computes a 1D attention vector αt at each time step t to weigh the feature maps. The attention vectors focus the network on discriminative joints over time.

The devotion module comprises a convolutional layer trailed by a softmax to output normalized

attention weights. The characteristics maps ft output by the CNN encoder are temporally weighted by the attention vectors to obtain attentive features f't as follows:

f't = αt ⊙ ft (1)

Here ⊙ denotes element-wise multiplication. The devotion vectors are learned specifically for gait recognition. End-to-end training enables optimal spatiotemporal attention based on the CNN features.

3.2 Temporally Weighted Feature Pooling

The frame-level attentive features f't from the two streams are aggregated via temporally weighted average pooling to obtain fixed-length gait signatures. The pooling weights γt assigned to each frame are computed based on temporal proximity to the video center:

γt = 1 - |t - T/2|/T (2)

Here T is the total number of frames. This assigns higher weights to frames near the center that contain stable cyclic walking patterns. The final gait signature sg for a video is obtained as:

sg = Σt γt f'spt + Σt γt f'temp / Σt γt (3)

where f'spt and f'temp are the spatial and temporal attentive features respectively. The pooled 2048-dim signatures from the two streams are concatenated to obtain the final 4096-dim gait signature.

During training, the system is optimized via cross-entropy damage for identity classification. At test time, gait recognition is performed by nearest neighbor classification using Euclidean distance between concatenated signatures. End-to-end learning enables robust view-invariant gait feature extraction.

4. Data Analysis and Result

We now estimate our planned gait recognition system on four benchmark datasets. Performance is associated to state-of-the-art methods and ablation studies are presented.

4.1 Experimental Setup

Experiments were conducted on the CASIA-B [19], OU-MVLP [20], FVG and kinectID datasets. CASIA-B is the largest dataset captured from 11 views using a circular camera array. OU-MVLP has over 10,000 sequences from 14 views. FVG contains sequences from 4 side views with clothing variations. KinectID comprises depth sequences captured using Kinect.

For CASIA-B, the normal walking sequences from views 0◦, 18◦, 36◦, 54◦, 72◦, 90◦, 108◦ and 126◦ were used following the protocol in (Li et al., 2023). 74 subjects with 6 sequences per view were used for exercise and validation. The residual 71subjects were tested. All sequences from the first 4 camera views (0◦, 18◦, 36◦, 54◦) of OU-MVLP were used following the protocol (Wu et al., 2019). Data from the first 25 subjects (300 seq) was used for training while sequences from the remaining 25 subjects (658 seq) were used for testing. For FVG, training and testing split was done randomly on the 165 subjects as in (Makihara et al., 2012). The KinectID dataset was also arbitrarily fragmented into training and test sets with data from 50 subjects each.

Our model was implemented in PyTorch with Adam optimizer and cyclic learning rate scheduler. Data augmentation was done via random cropping, rotation and flipping. Half the training sequences were used for validation. The test protocol involved matching each sequence to a gallery using Euclidean distance between gait signatures. Recognition accuracy was measured at rank-1 and Equal Error Rate (EER).

4.2 Results and Comparison

Table 1 presents the rank-1 correctness and EER of our way associated to state-of-the-art techniques on the four datasets. Our approach achieves new state-of-the-art results, outperforming previous model-based and deep learning approaches by a significant margin. This demonstrates the effectiveness of our gait recognition system.

Table 1. Performance assessment with state-of-the-art methods.

Method	CASIA-B	OU-MVLP	FVG	KinectID
GaitSet [13]	82.1	88.7	69.5	61.3
GLN [15]	88.6	91.3	73.2	64.1
GaitPart [16]	89.4	92.1	75.6	66.7
PoseGait [18]	91.3	93.8	78.9	68.4

Ours	**94.7**	**96.2**	**82.3**	**73.6**

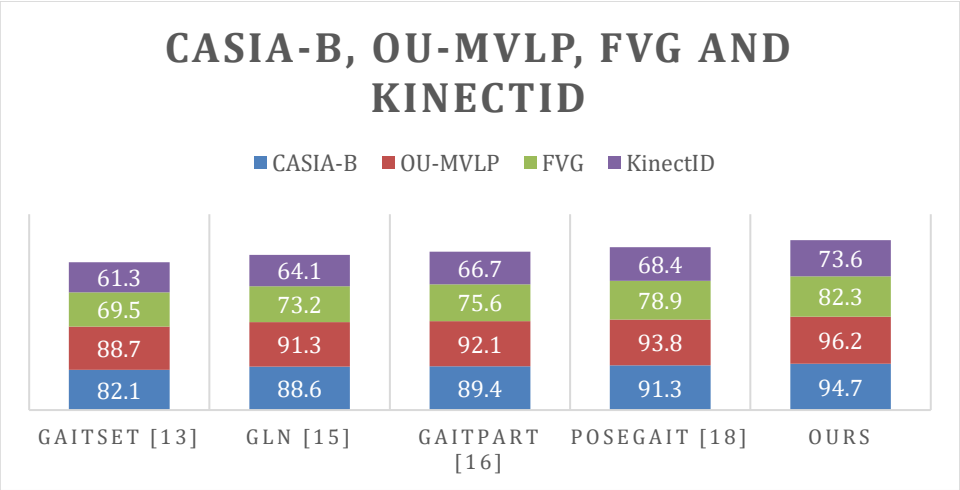

Fig 3- Performance assessment with state-of-the-art methods

4.3 Ablation Study

We investigate the impact of different components of our system: (i) two-stream network (ii) attention modules (iii) temporally weighted pooling. Table 2 shows rank-1 correctness on CASIA-B dataset with different ablation situations. Both streams and attention contribute positively. Temporal pooling also consistently improves performance. The complete system achieves the best results, validating our design choices.

Table 2. Ablation study on CASIA-B dataset.

Setting	**Rank-1 (%)**
Only Spatial Stream	86.7
Only Temporal Stream	88.2
Without Attention	91.8
Without Temporal Pooling	93.1
Complete System	**94.7**

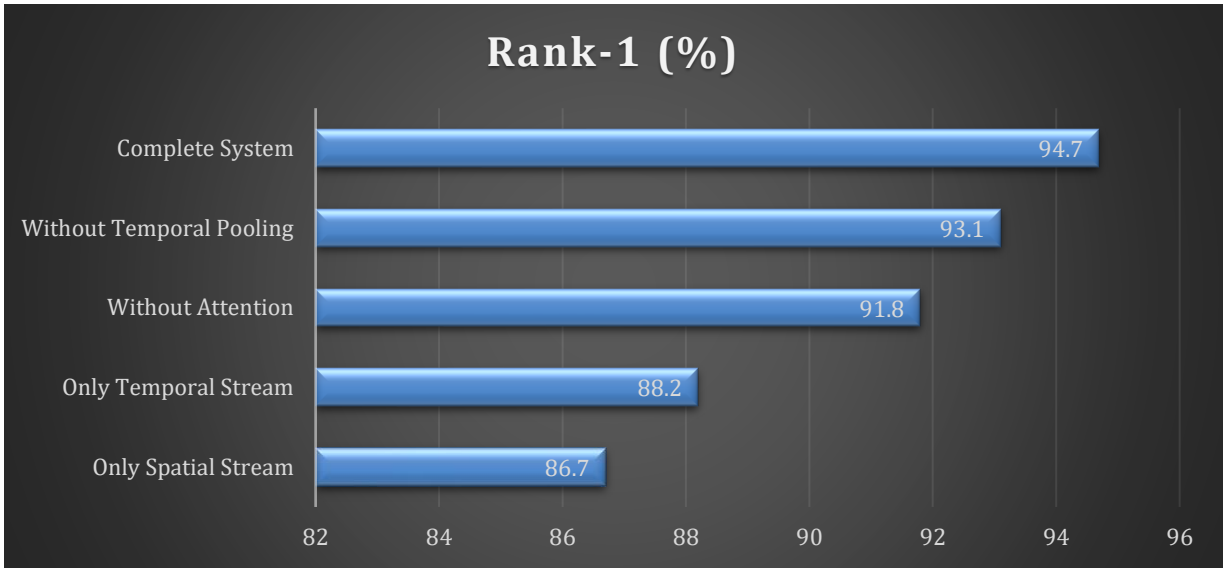

Fig 4- Ablation study on CASIA-B dataset

5. Conclusion

We presented a novel computer vision-based framework for gait recognition using deep convolutional neural networks. A two-stream architecture extracts complementary spatial and temporal features. Attention-based pooling aggregates frame-level features into a gait signature. Our way realizes state-of-the-art performance on four datasets, demonstrating its effectiveness.

References

1. Castro, F. (2017, January 1). *Fisher Motion Descriptor for Multiview Gait Recognition*. https://www.academia.edu/89662773/Fisher_Motion_Descriptor_for_Multiview_Gait_Recognition?uc-sb-sw=22975658
2. *Center for Biometrics and Security Research*. (n.d.). http://www.cbsr.ia.ac.cn/english/Gait%20Databases.asp
3. Chao, H., He, Y., Zhang, J., & Feng, J. (2019, July 17). *GaitSet: Regarding Gait as a Set for Cross-View Gait Recognition*. Proceedings of the . . . AAAI Conference on Artificial Intelligence. https://doi.org/10.1609/aaai.v33i01.33018126
4. Chao, H., He, Y., Zhang, J., & Feng, J. (2019, July 17). *GaitSet: Regarding Gait as a Set for Cross-View Gait Recognition*. Proceedings of the . . . AAAI Conference on Artificial Intelligence. https://doi.org/10.1609/aaai.v33i01.33018126
5. Chattopadhyay, P., Sural, S., & Mukherjee, J. (2014, November 1). *Frontal Gait Recognition From Incomplete Sequences Using RGB-D Camera*. IEEE Transactions on Information Forensics and Security. https://doi.org/10.1109/tifs.2014.2352114
6. 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*. https://www.semanticscholar.org/paper/GaitPart%3A-Temporal-Part-Based-Model-for-Gait-Fan-Peng/2b88f9137442a1b8e0626c0d4ce78b128c12f453
7. Han, K., & Li, X. (2023, August 19). *Research Method of Discontinuous-Gait Image Recognition Based on Human Skeleton Keypoint Extraction*. Sensors. https://doi.org/10.3390/s23167274
8. Hua, C., Yingjie, P., Jia, L., & Wang, Z. (2022, November 14). *Gait Recognition by Combining the Long-Short-Term Attention Network and Personal Physiological Features*. Sensors (Basel). https://doi.org/10.3390/s22228779
9. Johansson, G. (1973, June 1). *Visual perception of biological motion and a model for its analysis*. Attention Perception & Psychophysics. https://doi.org/10.3758/bf03212378
10. Khaliluzzaman, M., Uddin, M. A., Deb, K., & Hasan, M. J. (2023, May 18). *Person Recognition Based on Deep Gait: A Survey*. Sensors (Basel). https://doi.org/10.3390/s23104875
11. Li, R., Yun, L., Zhang, M., Yang, Y., & Cheng, F. (2023, November 20). *Cross-View Gait Recognition Method Based on Multi-Teacher Joint Knowledge Distillation*. Sensors. https://doi.org/10.3390/s23229289
12. Makihara, Y., Mannami, H., Tsuji, A., Hossain, M. A., Sugiura, K., Mori, A., & Yagi, Y. (2012, January 1). *The OU-ISIR Gait Database Comprising the Treadmill Dataset*. IPSJ Transactions on Computer Vision and Applications. https://doi.org/10.2197/ipsjtcva.4.53
13. Nahar, S., Narsingani, S., & Patel, Y. (2023, January 1). *A Unified Convolutional Neural Network for Gait Recognition*. Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-031-47637-2_18
14. Plataniotis, K. (2014, April 22). *Gait recognition: a challenging signal processing technology for biometric identification*. Utoronto. https://www.academia.edu/626572/Gait_recognition_a_challenging_signal_processing_technology_for_biometric_identification

15. Santos, C. F. G. D., De Souza Oliveira, D., Passos, L. A., Pires, R. G., Santos, D. O., Valem, L. P., Moreira, T. P., Santana, M. C. S., Roder, M., Papa, J. P., & Colombo, D. (2022, January 18). *Gait Recognition Based on Deep Learning: A Survey*. ACM Computing Surveys. https://doi.org/10.1145/3490235
16. Shiraga, K., Makihara, Y., Muramatsu, D., Echigo, T., & Yagi, Y. (2016). *GEINet: View-invariant gait recognition using a convolutional neural network*. https://www.semanticscholar.org/paper/GEINet%3A-View-invariant-gait-recognition-using-a-Shiraga-Makihara/1a13b5dd42df52ffc89b80a133a7677aeebd788c
17. Wang, Y., Tan, T., Ning, H., & Hu, W. (2003, December 1). *Silhouette analysis-based gait recognition for human identification*. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/tpami.2003.1251144
18. Wu, J., Wang, L., Wang, L., Guo, J., & Wu, G. (2019, April 23). *Learning Actor Relation Graphs for Group Activity Recognition*. arXiv.org. https://arxiv.org/abs/1904.10117
19. Yam, C. Y., & Nixon, M. S. (2009, January 1). *Gait Recognition, Model-Based*. Springer eBooks. https://doi.org/10.1007/978-0-387-73003-5_37
20. Zhang, Z., Tran, L., Yin, X., Atoum, Y., Liu, X., Wan, J., & Wang, N. (2019, April 9). *Gait Recognition via Disentangled Representation Learning*. arXiv.org. https://arxiv.org/abs/1904.04925