BPMT: Body Part as Modality Transformer for Efficient and Accurate Gait Recognition

Yusen Peng
The Ohio State University
Columbus, OH, USA
peng.1007@buckeyemail.osu.edu

Alper Yilmaz
The Ohio State University
Columbus, OH, USA
yilmaz.15@osu.edu

Abstract

Gait recognition is one of the most common applications in computer vision, which has been playing a crucial role in healthcare, crime analysis, and personal identification in general. In the era of deep learning, gait recognition has been empowered by neural networks and foundation models to extract meaningful feature representations. In this paper, the lightweight RTMPose model is integrated to efficiently extract 2D human poses, and a custom transformer is designed to perform accurate gait recognition upon the extracted human poses. Our research goal is to design and implement BPMT, Body Part as Modality Transformer, which treats individual body parts as different modalities to fit into a two-stage pretraining framework. We aim to outperform state-of-the-art models in terms of recognition accuracy and runtime complexity in the meantime. The code is in the GitHub Repository: https://github./Yusen-Peng/BPMT.

Keywords: Gait Recognition, Pose Estimation, Vision Transformer, Computer Vision

1 Introduction

Gait recognition, known as identifying people based on the way they walk, commonly gathered from video files [1], has multiple distinct advantages as a biometric [2]: human gait is characterized by the flexibility of being captured without physical proximity, and the permissiveness of low resolution [2]. A large number of human gait reference datasets have become available recently: a full-body gait dataset of able-bodied adults and stroke survivors has been collected to observe typical gait and aging effects [3]; another gait dataset that has combined anthropometric traits with gait parameters was acquired to facilitate patient evaluation of their physical status for health professionals [4]. Gait biometrics have been extensively studied under a standardized gait recognition pipeline: data acquisition, feature representation, dimension reduction, and classification [2]. To acquire human gait data, four devices have been widely employed: camera; accelerometer; floor sensor; continuous wave radar [2]. Feature representation can be classified into 2D representations and 3D representations based on the number of dimensions [1], or be divided into model-based representations and model-free representations depending on whether an explicit human body is defined [1], [2]. Both model-based and model-free approaches have been well studied [1]: a set of articulated 3D point cylinders was proposed as an explicit 3D gait model with both structural features and dynamic features [5]; on the other hand, GaitSet, a model-free method, was designed to learn human gait by viewing it as a set of independent, permutation-safe silhouettes from distinct videos [6]. Dimension reduction can be classified into feature reduction and outlier removal [2]: MPCA combined with LDA has been widely used for feature reduction to filter out irrelevant features [7]; Gaussian Mixture Model has been used to partition an image into hip, knee, ankle, and outlier [8]. The outlier region shall be removed to minimize noisy signals [2]. For classification, a variety of models have been studied and utilized including Hidden Markov Model (HMM) [9], Convolutional Neural Networks (CNNs) [10], and K Nearest Neighbors (KNN) [5].

2 Related Work

2.1 Pose Estimation

Pose estimation algorithms are classified into two modes: top-down approach and bottom-up approach [11]. With the overhead of applying an object detector to crop a human bounding box before employing the pose estimator, the top-down approach requires minimal input image res-

olutions while preserving decent speed and accuracy [11]; on the other hand, pinpointing instance-independent keypoints directly from a given image, the bottom-up approach suffers from strict input resolution to group identified keypoints into actual human poses [11]. The estimation of poses can be reduced to different tasks: heatmap-based methods, which rely heavily on the transformation of the keypoint format between the original keypoint coordinates and the heatmaps, outperform estimates directly from the keypoint coordinates without any transformation [12]; additionally, regression-based methods are characterized by efficient computation and permissive input resolution [12]; in addition, SimCC, a classification-based method, discretizes both axes into uniform bins to locate keypoints in a bin-wise manner [13].

2.2 Gait Recognition Model

Other than statistical models, deep learning has shown substantial potential in the classification stage of gait recognition, with a variety of different neural network architectures [14]. The most widely used architecture is convolutional neural networks (CNNs), which learn body shape embeddings by generating feature maps [14]. CNNs, often combined with other neural network architectures including autoencoders and LSTMs, generally do not require complex deep layers while still preserving sufficient encoding of gait frames [14]. Another prevalent architecture, recurrent neural networks (RNNs), can be applied in gait recognition in three distinct ways: directly learning temporal relationships of joint coordinates; combining with CNNs to learn both spatial and temporal features; learning partial representations within the same gait frame [14]. Based on the transformer architecture, deep auto-encoder (DAE) extracts bottleneck features into a latent space and minimizes the reconstruction error while decoding bottleneck features [14]. Generative Adversarial Networks (GANs), usually characterized by a generator and a discriminator, are robust to gait variations including viewpoints and clothing [14]. The generator is responsible for mixing fake samples into real ones, while the discriminator aims to distinguish between real samples and fake samples [14].

2.3 Transformer Architecture

In spite of the dominance of convolutional neural networks in the field of gait recognition, transformer architectures with an attention mechanism are able to capture and emphasize features in the specific regions of an image [15]. Most transformer-based models are conceptually similar, but they can be distinguished by different ways to learn embeddings and apply attention. Different types of embeddings have been adopted

in a variety of transformer architectures. Gait-ViT starts with partitioning gait silhouettes into fragments called patches, which will be concatenated into a time series [15]; then Gait-ViT learns patch embeddings along with positional embeddings, which are crucial to preserving the structural information within the original image [15]. On the other hand, instead of using silhouettes, IIP-Transformer partitions *human skeletons* into exactly 5 parts to extract patch embeddings [16]. For attention, in Gait-ViT, both patch embeddings and positional embeddings are passed into a transformer encoder, which consists of multiple transformer blocks with the self-attention mechanism [15]. However, a large number of transformer architectures are not limited to traditional self-attention: IIP-Transformer, with a total of 5 body parts defined and spatially partitioned, proposes Intra-Inter-Part self-attention (IIPA) to aggregate both joint-level and part-level tokens and sum up their attention outputs accordingly [16].

3 Research Motivation

Although a massive number of gait recognition models have been proposed, and multiple architectures including IIP-Transformer [16], STSA-Net [17], and IGFormer [18] have experimented with dividing the body poses explicitly into multiple meaningful parts or segments, no existing work has attempted to integrate body-part-aware transformers into the two-phase masked pretraining framework proposed in OmniVec2 [19], which was originally designed to learn multi-modal representations to perform multiple tasks. In this work, we attempt to address the following problem: can we treat different human parts as different modalities to integrate body-part-aware transformers into the two-phase masked pretraining framework proposed in OmniVec2 [19]?

4 Methodology

4.1 BPMT: Pipeline

BPMT pipeline is shown in Figure 1. During training, 2D human pose data from Gait3D are fed into BPMT. During inference, a real-time video can be fed into the RTMPose module to efficiently extract pose estimation, which will be the input to the trained BPMT.

4.2 BPMT: Architecture

BPMT architecture is shown in Figure 2. Estimated keypoints are divided into 5 parts and placed in the modality pool. During the first masked pre-training phase, each part as a modality is passed into the IIP-Transformer [16]; during the second pre-training phase, one pair of modality features learned from the transformer

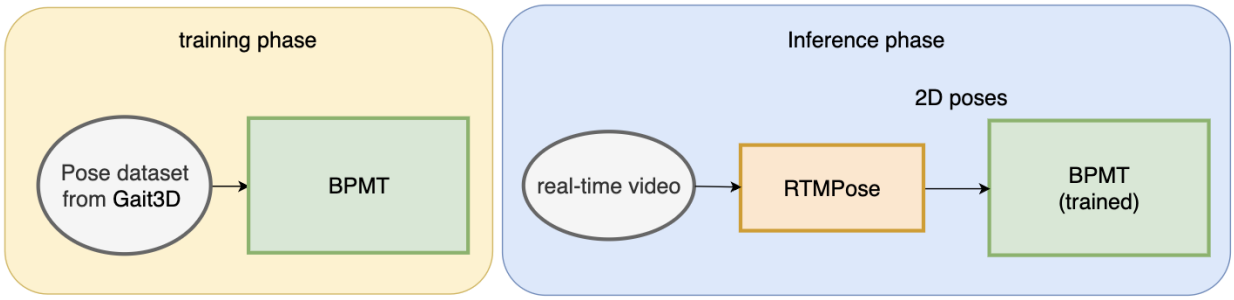

Figure 1: BPMT pipeline.

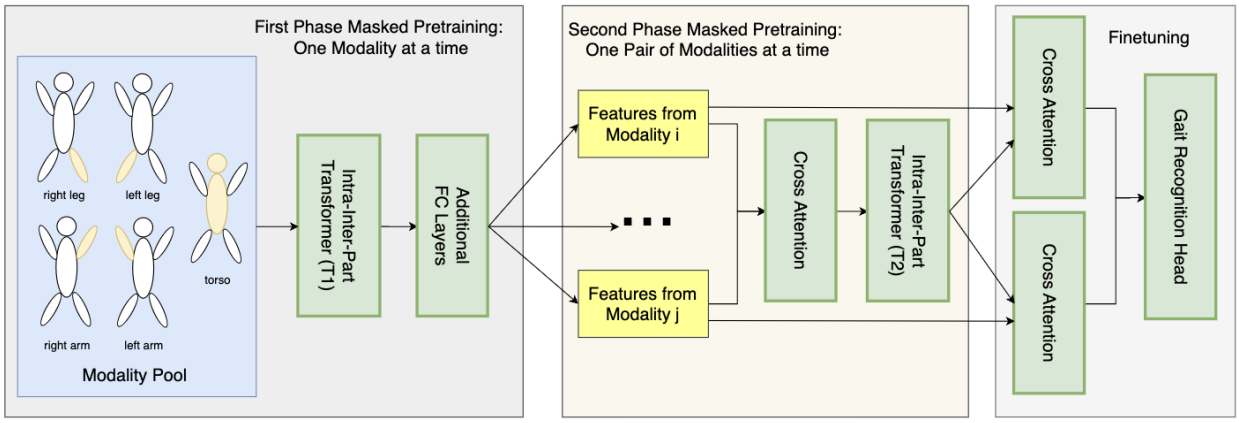

Figure 2: BPMT architecture.

is passed into a cross-attention layer, followed by another IIP-Transformer [16]. During fine-tuning, the output from the second transformer is combined with modality features learned from the first transformer, both of which are passed into another cross-attention layer, followed by the final task-specific head: gait recognition.

4.3 Training Dataset: Gait3D

Gait3D, a large-scale gait recognition dataset consisting of thousands of subjects under an unconstrained environment, gathers 3D meshes and 3D skeletons in addition to traditional 2D poses and 2D silhouettes. Cross-camera tracking has been employed to effectively cluster sequences of the same subject across all different cameras [1]. To be able to address the issue where two people can be highly overlapped, making it challenging for person tracking, human annotators were hired to clean up such overlapping sequences [1]. In spite of various rich representations including 3D meshes provided by Gait3D [1], to our interest, only 2D pose/skeleton data is collected and utilized for training purposes.

4.4 Pose Estimation: RTMPose

RTMPose, a high-performance pose estimation framework, is used for 2D pose feature representations in our work [11]. RTMPose pipeline starts with RTMDet [20], a computationally efficient, suitable, and available real-time bounding box detector, which has eliminated the inference running time bottleneck to separate each person within a single image [11]. CSPNeXt, a neural network that can take advantage of large kernel sizes to extract high-frequency features and use global average pooling for low-frequency feature extraction [21], was chosen as the backbone of RTMPose for its superior balance between efficiency and precision, with additional training techniques including proper pre-training and strengthening and then weak two-stage training enhancement [11]. SimCC [13], as a lightweight yet effective pipeline, was used in RTMPose to estimate keypoints within a classification framework enhanced with a larger convolution kernel and self-attention mechanism [11]. In terms of the inference pipeline, the real-time friendly skip frame detection mechanism [22], in which detection is only performed on every K frames, was used to boost efficiency while still preserving competitive performance [11]. rtmlib, a super lightweight library to conduct pose estimation based on RTMPose models, is employed to derive 2D human pose representations during the inference phase [23].

5 Project Timeline

The timeline for this research project is in the table below. The already finished tasks are marked as teal; the tasks working in progress are marked as blue; the future tasks are marked as gray.

Tasks	**Timeline**
Lit. review: gait recognition	Feb
Lit. review: pose estimation	Feb
experiment: RTMPose	Feb
experiment: Gait3D dataset	Feb
Lit. review: NN models	Feb
Lit. review: transformers	Feb
Writing: proposal	Feb to Mar
Design: transformer architecture	Mar to July
Coding: transformer architecture	Mar to July
experiment: training	May to July
experiment: inference	Jun to Aug
experiment: comparative study	Aug to Sept
writing: thesis	May to Oct
Coding: GitHub documentation	Oct
submit the paper to ICLR	by Oct
submit the paper to CVPR	by Nov
submit the paper to ICML	by Jan (next year)
submit the paper to ICCV	by Mar (next year)
4999: 2*3 credit hours	Aug to Apr (next year)
submit the paper to NeurIPS	by May (next year)

6 Personal Statement

I have had a solid research background in the field of applied machine learning with more than one year of experience in time series analysis. I am familiar with various machine learning algorithms, neural network architectures, and am comfortable with sklearn and PyTorch programming. However, I am relatively new to the field of computer vision and not familiar with gait recognition; thus, I have to do an extensive literature review on the subject of gait recognition. This project will prepare me for my future academic career in applied machine learning with the most fundamental skills of conducting an independent study, such as doing a literature review, identifying existing research gaps, designing transformer architectures, and implementing an end-to-end machine learning pipeline. In this project, in order to develop and train BPMT, I will be using the server from the Photogrammetric Computer Vision Lab directed by Dr. Alper Yilmaz.

References

[1] J. Zheng, C. Chen, S. Wang, Y. Yang, Q. Wu, and L. Shen, "Gait recognition in the wild with dense 3d representations and a benchmark," *arXiv preprint arXiv:2204.02569*, 2022. [Online]. Available: https://arxiv.org/abs/2204.02569.

[2] C. Wan, L. Wang, and V. V. Phoha, "A survey on gait recognition," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–35, 2018. DOI: 10.1145/3230633.

[3] T. V. Criekinge *et al.*, "A full-body motion capture gait dataset of 138 able-bodied adults across the life span and 50 stroke survivors," *Scientific Data*, vol. 10, no. 1, 2023. DOI: 10.1038/s41597-023-02767-y.

[4] J. Zafra-Palma, N. Marín-Jiménez, J. Castro-Piñero, *et al.*, "Health & gait: A dataset for gait-based analysis," *Scientific Data*, vol. 12, no. 44, 2025. DOI: 10.1038/s41597-024-04327-4.

[5] G. Ariyanto and M. S. Nixon, "Model-based 3d gait biometrics," in *Proceedings of the International Joint Conference on Biometrics (IJCB)*, Washington, DC, USA, 2011, pp. 1–7.

[6] H. Chao, Y. He, J. Zhang, and J. Feng, "Gaitset: Regarding gait as a set for cross-view gait recognition," *arXiv preprint arXiv:1811.06186*, 2018. [Online]. Available: https://arxiv.org/abs/1811.06186.

[7] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Boosting lda with regularization on mpca features for gait recognition," in *2007 Biometrics Symposium*, Baltimore,

MD, USA, 2007, pp. 1–6. DOI: 10.1109/BCC.2007.4430542.

[8] E. F. A. Mashagba, F. F. A. Mashagba, and M. O. Nassar, "Simple and efficient marker-based approach in human gait analysis using gaussian mixture model," *Australian Journal of Basic and Applied Sciences*, vol. 8, no. 1, pp. 137–147, 2014.

[9] A. Kale, A. N. Rajagopalan, N. Cuntoor, and V. Kruger, "Gait-based recognition of humans using continuous hmms," in *Proceedings of Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, Washington, DC, USA, 2002, pp. 336–341. DOI: 10.1109/AFGR.2002.1004176.

[10] M. Alotaibi and A. Mahmood, "Improved gait recognition based on specialized deep convolutional neural networks," in *2015 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, Washington, DC, USA, 2015, pp. 1–7. DOI: 10.1109/AIPR.2015.7444550.

[11] T. Jiang *et al.*, "Rtmpose: Real-time multi-person pose estimation based on mmpose," *arXiv.org*, 2023, Accessed: Feb. 10, 2025. [Online]. Available: https://arxiv.org/abs/2303.07399.

[12] J. Li, C. Lan, W. Zeng, Y. Zhang, L. Zhang, and Z. Liu, "Human pose regression with residual log-likelihood estimation," *arXiv*, vol. abs/2107.11291, 2021, Accessed: Feb. 10, 2025. arXiv: 2107.11291 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2107.11291.

[13] J. Huang, Z. Zhu, F. Guo, G. Huang, and D. Du, "The devil is in the details: Delving into unbiased data processing for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Accessed: Feb. 10, 2025, 2020, pp. 5700–5709. [Online]. Available: https://arxiv.org/abs/1911.07524.

[14] A. Sepas-Moghaddam and A. Etemad, "Deep gait recognition: A survey," *CoRR*, vol. abs/2102.09546, 2021. arXiv: 2102.09546. [Online]. Available: https://arxiv.org/abs/2102.09546.

[15] J. N. Mogan, C.-P. Lee, K. Lim, and K. Anbananthen, "Gait-vit: Gait recognition with vision transformer," *Sensors (Basel, Switzerland)*, vol. 22, Sep. 2022. DOI: 10.3390/s22197362.

[16] Q. Wang, J. Peng, S. Shi, T. Liu, J. He, and R. Weng, "Iip-transformer: Intra-inter-part transformer for skeleton-based action recognition," *CoRR*, vol. abs/2110.13385, 2021. arXiv: 2110.13385. [Online]. Available: https://arxiv.org/abs/2110.13385.

[17] H. Qiu, B. Hou, B. Ren, and X. Zhang, "Spatio-temporal segments attention for skeleton-based action recognition," *Neurocomput.*, vol. 518, no. C, pp. 30–38, Jan. 2023, ISSN: 0925-2312. DOI: 10.1016/j.neucom.2022.10.084. [Online]. Available: https://doi.org/10.1016/j.neucom.2022.10.084.

[18] Y. Pang, Q. Ke, H. Rahmani, J. Bailey, and J. Liu, *Igformer: Interaction graph transformer for skeleton-based human interaction recognition*, 2022. arXiv: 2207.12100 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2207.12100.

[19] S. Srivastava and G. Sharma, "Omnivec2 - a novel transformer based network for large scale multimodal and multitask learning," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 27 402–27 414. DOI: 10.1109/CVPR52733.2024.02588.

[20] C. Lyu, W. Zhang, H. Huang, *et al.*, *Rtmdet: An empirical study of designing real-time object detectors*, 2022. arXiv: 2212.07784 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2212.07784.

[21] X. Chen, C. Yang, J. Mo, *et al.*, "CSP-NeXt: A new efficient token hybrid backbone," *Engineering Applications of Artificial Intelligence*, vol. 132, p. 107 886, 2024. DOI: 10.1016/j.engappai.2024.107886.

[22] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "Blazepose: On-device real-time body pose tracking," *CoRR*, vol. abs/2006.10204, 2020. arXiv: 2006.10204. [Online]. Available: https://arxiv.org/abs/2006.10204.

[23] T. Jiang, *Rtmlib*, https://github.com/Tau-J/rtmlib, 2023.