Hierarchical Spatio-Temporal Representation Learning for Gait Recognition (*Supplementary Material*)

Lei Wang[1,2], Bo Liu[1,2]*, Fangfang Liang[1,2], Bincheng Wang[1,2]
[1] College of Information Science and Technology, Hebei Agricultural University, China
[2] Hebei Key Laboratory of Agricultural Big Data, China
{20212060107, 20212060108}@pgs.hebau.edu.cn, {boliu, liangfangfang}@hebau.edu.cn

A. Supplementary Material

The supplementary material includes:

1. The detailed architecture of HSTL for the OUMVLP [1], GREW [4] and Gait3D [3] datasets.
2. Ablation experiment on the number of strips in hierarchical feature mapping.
3. Detailed partition settings for ARME.
4. Ablation experiment on the dimensions of applied FC in FTA.

A.1. Architectural Details of the HSTL framework on the OUMVLP, GREW, and Gait3D Datasets

In Table 2 of the paper, we have outlined the architecture of the proposed HSTL method for the CASIA-B dataset. In this section, we provide further details on the architecture of HSTL for the OUMVLP, GREW, and Gait3D datasets. These datasets contain larger numbers of subjects and more complex distributions compared to the CASIA-B dataset. For instance, the OUMVLP dataset only contains normal walking (NM) sequences, but it contains a significant age range among its subjects. Additionally, both GREW and Gait3D, collected in uncontrolled environments, have more complex conditions such as body occlusion and pose misalignment compared to CASIA-B and OUMVLP. HSTL uses unsupervised hierarchical clustering to extract hierarchical motion relationships for gait sequences, which guides its architectural design. To account for the complexity of these large datasets, we deepen the network by incorporating an additional level of hierarchical clustering. The resulting architecture design, as shown in Tables 1-3, enhances the adaptability of HSTL based on the different hierarchy results.

Table 1. The detailed architecture of the proposed HSTL on OUMVLP. The first column denotes the levels of the gait hierarchy and K_l is the number of groups at level l. C_{in} and C_{out} represent the input channel and output channel of each layer respectively. The body parts are indexed in spatial order from top to bottom, numbered 1 to 8.

Level	Block	Layer	C_{in}	C_{out}	Kernel	K_l	Parts Grouping
1	ARME	Conv3d	1	32	(3,3,3)	1	$\{\{1,2,3,4,5,6,7,8\}\}$
	ASTP						
2	ARME	Conv3d	32	32	(3,3,3)	3	$\{\{1\},\{2,3,4,5,6,7\},\{8\}\}$
		Conv3d	32	64	(3,3,3)		
	ASTP						
2	FTA	MaxPool	64	64	(3,1,1)	3	$\{\{1\},\{2,3,4,5,6,7\},\{8\}\}$
					(5,1,1)		
	ASTP						
3	ARME	Conv3d	64	128	(3,3,3)	5	$\{\{1\},\{2\},\{3,4,5\},\{6,7\},\{8\}\}$
		Conv3d	128	128	(3,3,3)		
	ASTP						
4	ARME	Conv3d	64	128	(3,3,3)	6	$\{\{1\},\{2\},\{3,4,5\},\{6\},\{7\},\{8\}\}$
		Conv3d	128	128	(3,3,3)		
	ASTP						
5	ASTP					8	$\{\{1\},\{2\},\{3\},\{4\},$ $\{5\},\{6\},\{7\},\{8\}\}$

A.2. Ablation Study on Different Strips Numbers

In HSTL, multiple adaptive spatio-temporal pooling (ASTP) modules are used to produce hierarchical gait embeddings. The effect of varying levels of ASTP modules on performance is shown in Table 4. The results indicate that as the number of gait embeddings increases, the accuracy improves across different walking conditions [2].

A.3. Detailed Setting of the ARME

The detailed partitioning of each setting in Figure 5 of the paper is shown in Table 5.

A.4. Ablation Study on Different Dimensions for FC

In FTA, the FC layer targets the $C \times \frac{T}{3}$ dimension to aggregate temporal information at multiple scales, guiding our choice of this dimension. In addition, we tested the FC layer applied to different dimensions ($H \times W$, C, and

*Corresponding Author

Table 2. The detailed architecture of the proposed HSTL on GREW. The first column denotes the levels of the gait hierarchy and K_l is the number of groups at level l. C_{in} and C_{out} represent the input channel and output channel of each layer respectively. The body parts are indexed in spatial order from top to bottom, numbered 1 to 8.

Level	Block	Layer	C_{in}	C_{out}	Kernel	K_l	Parts Grouping
1	ARME	Conv3d	1	32	(3,3,3)	1	$\{\{1,2,3,4,5,6,7,8\}\}$
	ASTP						
2	ARME	Conv3d	32	32	(3,3,3)	2	$\{\{1,2\},\{3,4,5,6,7,8\}\}$
		Conv3d	32	64	(3,3,3)		
	ASTP						
2	FTA	MaxPool	64	64	(3,1,1)	2	$\{\{1,2\},\{3,4,5,6,7,8\}\}$
					(5,1,1)		
	ASTP						
3	ARME	Conv3d	64	128	(3,3,3)	3	$\{\{1\},\{2\},\{3,4,5,6,7,8\}\}$
		Conv3d	128	128	(3,3,3)		
	ASTP						
4	ARME	Conv3d	64	128	(3,3,3)	4	$\{\{1\},\{2\},\{3,4,5,6,7\},\{8\}\}$
		Conv3d	128	128	(3,3,3)		
	ASTP						
5	ASTP					8	$\{\{1\},\{2\},\{3\},\{4\},$ $\{5\},\{6\},\{7\},\{8\}\}$

Table 3. The detailed architecture of the proposed HSTL on Gait3D. The first column denotes the levels of the gait hierarchy and K_l is the number of groups at level l. C_{in} and C_{out} represent the input channel and output channel of each layer respectively. The body parts are indexed in spatial order from top to bottom, numbered 1 to 8.

Level	Block	Layer	C_{in}	C_{out}	Kernel	K_l	Parts Grouping
1	ARME	Conv3d	1	32	(3,3,3)	1	$\{\{1,2,3,4,5,6,7,8\}\}$
	ASTP						
2	ARME	Conv3d	32	32	(3,3,3)	2	$\{\{1,2,3,4,5\},\{6,7,8\}\}$
		Conv3d	32	64	(3,3,3)		
	ASTP						
2	FTA	MaxPool	64	64	(3,1,1)	2	$\{\{1,2,3,4,5\},\{6,7,8\}\}$
					(5,1,1)		
	ASTP						
3	ARME	Conv3d	64	128	(3,3,3)	4	$\{\{1\},\{2,3,4,5\},\{6,7\},\{8\}\}$
		Conv3d	128	128	(3,3,3)		
	ASTP						
4	ARME	Conv3d	64	128	(3,3,3)	5	$\{\{1\},\{2\},\{3,4,5\},\{6,7\},\{8\}\}$
		Conv3d	128	128	(3,3,3)		
	ASTP						
5	ASTP					8	$\{\{1\},\{2\},\{3\},\{4\},$ $\{5\},\{6\},\{7\},\{8\}\}$

Table 4. Ablation study investigates the effectiveness of varying the number of ASTP modules on the average rank-1 accuracy on the CASIA-B dataset. (ASTP$^{(*)}$ denotes the ASTP module associated with a specific hierarchical level.)

ASTP$^{(1)}$	ASTP$^{(2)}$	ASTP$^{(2)}$	ASTP$^{(3)}$	ASTP$^{(4)}$	NM	BG	CL	Mean
				✓	97.5	95.1	86.8	93.1
			✓	✓	97.6	95.3	87.7	93.5
		✓	✓	✓	97.7	95.6	88.4	93.9
	✓	✓	✓	✓	97.9	95.8	88.7	94.1
✓	✓	✓	✓	✓	**98.1**	**95.9**	**88.9**	**94.3**

$\frac{T}{3}$). Results show that applying the FC layer to the $C \times \frac{T}{3}$ dimension yields the best performance, as is shown in

Table 5. Detailed group settings for ARME ablation studies.

Setting	Parts Grouping		
	ARME$^{(1)}$	ARME$^{(2)}$	ARME$^{(3)}$
1-1-1	$\{\{1,2,3,4,5,6,7,8\}\}$	$\{\{1,2,3,4,5,6,7,8\}\}$	$\{\{1,2,3,4,5,6,7,8\}\}$
2-2-2	$\{\{1,2,3,4\},\{5,6,7,8\}\}$	$\{\{1,2,3,4\},\{5,6,7,8\}\}$	$\{\{1,2,3,4\},\{5,6,7,8\}\}$
4-4-4	$\{\{1,2\},\{3,4\},\{5,6\},\{7,8\}\}$	$\{\{1,2\},\{3,4\},\{5,6\},\{7,8\}\}$	$\{\{1,2\},\{3,4\},\{5,6\},\{7,8\}\}$
1-2-4	$\{\{1,2,3,4,5,6,7,8\}\}$	$\{\{1,2,3,4\},\{5,6,7,8\}\}$	$\{\{1,2\},\{3,4\},\{5,6\},\{7,8\}\}$
1-2-4*	$\{\{1,2,3,4,5,6,7,8\}\}$	$\{\{1,2,3,4,5\},\{6,7,8\}\}$	$\{\{1\},\{2,3,4,5\},\{6,7\},\{8\}\}$

Table 6.

Table 6. Ablation study examines the applied dimensions of the fully connected (FC) layers of the FTA modules in terms of average rank-1 accuracy on the CASIA-B dataset.

Settings (CASIA-B)	NM	BG	CL	Mean
FTA ($H \times W$)	97.9	95.6	88.2	93.9
FTA (C)	97.6	95.3	88.7	93.9
FTA ($\frac{T}{3}$)	97.5	95.7	**88.9**	94.0
FTA ($\frac{T}{3} \times C$)	**98.1**	**95.9**	**88.9**	**94.3**

References

[1] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ TCVA*, 10:1–14, 2018. 1

[2] Lei Wang, Bo Liu, Bincheng Wang, and Fuqiang Yu. Gaitmm: Multi-granularity motion sequence learning for gait recognition, 2022. 1

[3] Jinkai Zheng, Xinchen Liu, Wu Liu, Lingxiao He, Chenggang Yan, and Tao Mei. Gait recognition in the wild with dense 3d representations and a benchmark. In *CVPR*, pages 20228–20237, 2022. 1

[4] Zheng Zhu, Xianda Guo, Tian Yang, Junjie Huang, Jiankang Deng, Guan Huang, Dalong Du, Jiwen Lu, and Jie Zhou. Gait recognition in the wild: A benchmark. In *ICCV*, pages 14789–14799, 2021. 1