OpenGait: Revisiting Gait Recognition Toward Better Practicality

Chao Fan[1,2], Junhao Liang[1,2], Chuanfu Shen[3,1], Saihui Hou[4,5],
Yongzhen Huang[4,5], Shiqi Yu[1,2*]

[1] Department of Computer Science and Engineering, Southern University of Science and Technology
[2] Research Institute of Trustworthy Autonomous System, Southern University of Science and Technology
[3] Department of Industrial and Manufacturing Systems Engineering, The University of Hong Kong
[4] School of Artificial Intelligence, Beijing Normal University [5] WATRIX.AI

{12131100, 12132342, 11950016}@mail.sustech.edu.cn, {housaihui, huangyongzhen}@bnu.edu.cn, yusq@sustech.edu.cn

8. Supplementary Material

The source code of GaitBase is avaliable at https://github.com/ShiqiYu/OpenGait. In this section, we first explore the effectiveness of several usual spatial data augmentation operations. Then, we conduct comprehensive experiments to analyze the effect of random training input length. Lastly, we talk about some future works that are worth further exploration.

8.1. Effect of Spatial Data Augmentation

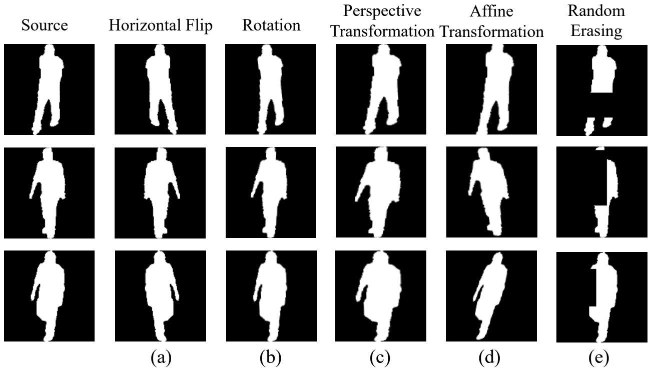

Figure 6. The visualization of source image with different spatial data augmentation operations. In Rotation, the twist angle is randomly sampled from $[-10°, +10°]$. In Perspective Transformation, the source axes are randomly skewed within 10 pixels to produce the transform axes of perspective. In Affine Transformation, we perform Rotation on source images and then shear them with their level ranging on $[-5 \times 10^{-3}, 5 \times 10^{-3}]$. In Random Erasing, we use the original hyper-parameters [49] for image classification. For each input sequence, the probability of performing spatial augmentation operations is set to 0.5.

Table 9. Ablation study for spatial data augmentation with the fixed length training input. Rank-1 accuracies (%) are reported on CASIA-B* and Gait3D, HF for Horizontal Flip, R for Rotation, PT for Perspective Transformation, AT for Affine Transformation, and RE for Random Erasing. The bracket indicates that the performance outperforms the GaitBase without data augmentation.

Group	Data Augmentation					CASIA-B*	Gait3D
	HF	R	PT	AT	RE		
-						86.8	54.7
(a)	✓					86.5	(59.5)
(b)		✓				(87.4)	(60.5)
(c)			✓			86.6	(58.9)
(d)				✓		79.0	(55.8)
(e)					✓	(87.8)	54.1
(f)		✓			✓	**88.7**	-
(g)	✓	✓	✓			-	**(62.4)**

As shown in Fig. 6, we perform various spatial augmentation techniques to enlarge the data space and avoid overfitting of the model. We conduct an ablation study on two commonly used indoor and outdoor datasets, *i.e.*, CASIA-B*[1] and Gait3D, to evaluate the efficacy of these approaches experimentally. The results are shown in Table 9.

Horizontal Flip can largely simulate a mirror transformation of the filming viewpoint. In (a), we observe that although it fails to improve the performance of CASIA-B*, it significantly improves accuracy on Gait3D. This can be explained by the fact that CASIA-B* is recorded in a laboratory environment using an all-sided camera array, while Gait3D is captured in real-world conditions with comparatively fewer viewpoint changes per subject.

From the experiment (b) in Table 9, Rotation technique slightly benefits both CASIA-B* and Gait3D.

Perspective Transformation aims to simulate the effects

[*] Corresponding Author

[1] The conclusions obtained from the experiments on CASIA-B and CASIA-B* are consistent. Here we only present the results on CASIA-B* for brevity.

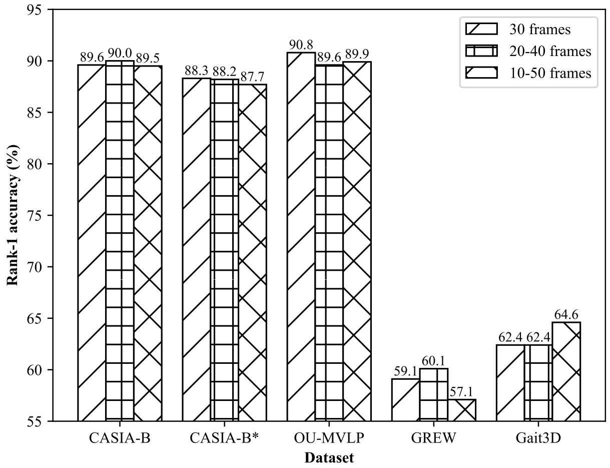

Figure 7. The effect of random training input length. 20-40 and 10-50 represent that the input lengths are uniformly distributed over 20 to 40 and 10 to 50.

of different camera heights. As shown in (c), our experiments indicate that this technique only has a significant impact on Gait3D dataset. The cause should be that there are no camera height changes in CASIA-B*, whereas there are such changes in Gait3D.

From the experiment (d) in Table 9, it appears that Affine Transformation is not able to effectively simulate the noisy factors present in both indoor CASIA-B* and outdoor Gait3D datasets, thereby failing to bring any performance gain on these datasets.

The main goal of Random Erasing [49] is to simulate the occlusion conditions and avoid the over-fitting problem in the spatial dimension. From (e), we note that the erasing operation is challenging to simulate the practical occlusion cases and thus makes almost no difference on Gait3D.

Building upon the above experimental results, we apply the Rotation and Random Erasing for the indoor datasets such as CASIA-B*, CASIA-B and OU-MVLP. On the other hand, for outdoor datasets like Gait3D and GREW, we employ the combination of Horizontal Flip, Rotation, and Perspective Transformation as the augmentation strategy. As evident from (f) and (g), it can be observed that the data augmentation approach brings accuracy improvements of 1.9% and 7.7% on CASIA-B* and Gait3D, respectively.

Based on the above analysis, we again expose the sizable gap between the experimental and practical gait data. Therefore, we propose that the research community should focus more attention on outdoor gait datasets for better practicality.

8.2. Effect of Random Training Input Length

In this subsection, we investigate the effect of random training input length on the final recognition performance. As shown in Fig. 7, we observe that the fixed length input works relatively optimal for the indoor dataset such as CASIA-B, CASIA-B* and OU-MVLP. On the other hand, the use of random length input yields superior performance for outdoor datasets, such as GREW and Gait3D. Theoretically, similar to the popular Dropout [45] technique, the usage of random sequence length can maintain consistency in features, regardless of the input length, thus easing the overfitting problem in the temporal dimension. In laboratory-acquired datasets, frame dropping and walking speed fluctuations are minimal, thereby resulting in a uniform number of frames in the gait cycle. As a result, the random training input length has little impact on indoor datasets such as CASIA-B, CASIA-B*, and OU-MVLP, which may be attributed to this consistent nature.

8.3. Future Work and Discussion

This paper presents a comprehensive benchmark study towards gait recognition applications, which includes a flexible codebase, a series of experimental reviews, and a robust baseline. In addition, here we highlight some subsequent works that are worth further exploration.

Gait Verification Task. The evaluation protocols of existing gait datasets mostly focus on identification (retrieval) tasks, resulting in gait verification scenarios being ignored in most cases. Typically, there are two categories of methods for performing verification processes: training a binary classifier [47, 48] or inferring a conclusive distance threshold to determine whether the two subjects come from the same identity. However, since clothing and viewpoint changes may dramatically impact gait appearance, reducing the intra-class distance is always a challenging issue for gait recognition. This poses a huge obstacle for gait verification applications. We encourage the research community to devote more attention to this complex topic, as it is widely needed for practical usage.

Stronger Baseline Model. Although the proposed baseline model, GaitBase, has achieved state-of-the-art performance on the largest outdoor gait dataset, GREW [50], with a rank-1 accuracy of 60.1%, there is still a significant gap in achieving an accurate enough gait recognition for real-world applications. Additionally, there has been a modeling shift from CNNs to Transformers [43, 44, 46] in many visual tasks. With its outstanding modeling capabilities, transformer-based gait recognition offers a fascinating solution to the challenges posed by outdoor environments, yet it has not received the attention it deserves. Therefore, the development of a stronger baseline model, such as a transformer-based model, remains a pressing issue for practical gait recognition.

Unsupervised Gait Recognition. The large-scale collection of annotated gait data in the wild is economically expensive and usually limited in the trade-off between the diversity and scale. For example, the largest outdoor gait dataset, GREW [50], covers over 20,000 subjects, but each subject, on average, only has about 4.5 walking sequences mostly captured from nearly front and back viewpoints. Additionally, it is challenging for outdoor gait datasets like GREW to include the long-term changes in clothing, age, hair, and body sizes for each subject as their collection process typically finishes in several months. Therefore, we consider learning the general gait representation, *i.e.*, prior identity knowledge, from unlabelled walking videos to be a challenging yet highly appealing task for further study.

Ethical Statements.

We are highly concerned about personal information security and argue that the improper use or abuse of gait recognition will threaten personal privacy. We believe that the development of vision techniques should only devote to the cause of human happiness.

Acknowledgment

We want to thank the reviewers for their efforts and the authors of the references for their inspiring insights and awesome achievements.

References

[43] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[44] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[45] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 2

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[47] Zifeng Wu, Yongzhen Huang, Liang Wang, Xiaogang Wang, and Tieniu Tan. 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, 2016. 2

[48] Kaihao Zhang, Wenhan Luo, Lin Ma, Wei Liu, and Hongdong Li. Learning joint gait representation via quintuplet loss minimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4700–4709, 2019. 2

[49] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1318–1327, 2017. 1, 2

[50] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14789–14799, 2021. 2, 3