Gait Transformer: Silhouette-Based View-Invariant Gait Recognition Model

Xuhang Chen and Chi-Man Pun, *Senior Member, IEEE*

I. Introduction

THE gait recognition is a significant classification task in biometric. Moreover, the gait data can be obtained remotely without human interference. Therefore, it plays an important role in video capture and human posture identification. Nowadays, surveillance cameras are irreplaceable in crime prevention and forensic evidence.

However, automatic and accurate gait recognition is far from completeness. A major reason is that there could be drastic deviation during human movement, such as view, clothing, different objects in company. These changes often impair classification greatly. Among all these factors, view is listed as the top because walking directions of subjects in actual scenario are unpredictable, which is the major problem we intend to solve in this work.

According to the survey[1], silhouettes is the most popular body representation of all time. Skeletons and hybrid representation are trending since they can enhance the accuracy efficiently. Taking popularity and real world application into consideration, we choose silhouettes solely as research data. In review of gait recognition methodology, apart from Convolutional Neural Network(CNN), Long Short-Term Memory(LSTM) and Generative Adversarial Network(GAN) are also common alternatives or ensemble parts. Depending on different types of silhouettes, the current state-of-the-art accuracy is 81.5% to 97.9%[1].

Recently, Vision Transformer(ViT)[2] is a popular architecture in computer vision field due to its self-attention mechanism. It can handle conventional recognition tasks and derive results comparable to the current methods such as CNN. Therefore, we suppose integrating ViT in gait recognition can increase accuracy and lead to remarkable performance.

II. METHODOLOGY

A. Vision Transformer

Transformer is formerly a very popular model in natural language processing(NLP) area[3]. It is a strong competitor to CNN and requires less computational power. The detailed representation of Vision Transformer can be found in Figure 1. Similar to NLP, images are also patched into fixed-size parts. Before feeding to encoder, some embeddings and tokens are mixed with the patches.

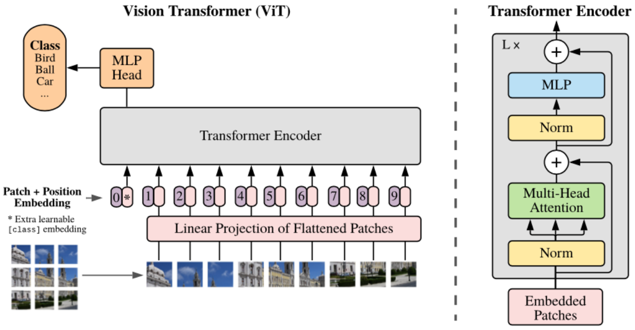

Fig. 1. Overview structure of ViT. All fixed-size patches are from the divided image. They are linearly placed with position embeddings. The resulting sequence of vectors are sent to a typical enconder in Transformer. To perform classification, the standard approach is adding an extra learnable ”classification token” to the sequence.

B. Loss Function

In our model, we select triplet margin loss(TM) and cross entropy loss(CE) as our loss functions. We use TM simply because there are three types of gaits. And we want our classifier to tell their difference. So except for the sample itself, we need another positive sample and a negative sample to enhance the difference. CE is a widely used loss function and it is the most popular loss in gait recognition task.

III. EXPERIMENT

A. Datasets

CASIA-B dataset[4] is a multi-view gait dataset. It contains 3 different conditions of gait, namely normal walking(NM), walking with a bag(BG), and walking with a coat(CL). And there are 11 views in total. It is the most widely used gait dataset so far.

B. Training Details

All image size is transformed to 224×224 pixels. The learning rate is 1e-4 using Adam optimizer and the weight decay is 1e-5.

C. Results and Analysis

The results of different models are displayed in Table 1. As we can see, our model performs good enough compared to the baseline traditional method. However, there is still room for improvement. We explore the process and list some potential factors for polishing and future enhancement.

The authors are with the Department of Computer and Information Science, University of Macau, Taipa, Macau (email: yc17491@umac.mo; cmpun@umac.mo).

1) Accuracy: Our model performs not as well as most state-of-the-art models. We think hyper-parameters, training configuration are main attributions.

2) Gait Variation: Although we try to eliminate the influence of view in classification. Walking condition is still a down factor for our work.

3) Training Time: The training time of ViT is relatively long compared to existing methods and models. In our experiments, we spend about 14 days for 100 epochs.

4) Transfer Learning: Since we spend enormous time training the ViT and still find it not very effective. We also find some other potential solution to facilitate the whole process. As [5] suggests, training ViT from the beginning is not a wise choice sometimes. Maybe we should apply transfer learning by using pretrained ViT models from other tasks.

TABLE I
STATE-OF-THE-ART RESULTS ON CASIA-B DATASET. NM, BG, AND CL MEAN NORMAL WALKING, WALKING WITH A BAG, AND WALKING WITH A COAT TEST PROTOCOLS.

Method		Performance			
Year	Name	NM	BG	CL	Average
2019	GaitNet[6]	93.9	82.6	63.2	79.9
2019	GaitSet[7]	95.0	87.2	70.4	84.2
2020	GaitPart[8]	96.2	91.5	78.7	88.8
2020	GLN[9]	96.8	94.0	77.5	89.4
2020	MT3D[10]	96.7	93.0	81.5	90.4
2021	Gait Transformer(Ours)	91.2	80.7	65.5	79.1

IV. CONCLUSION

In this article, we apply ViT in gait recognition, which is the first one in this area. We show that transformer is a potential successful model in this field and further research can explore an optimized ViT using transfer learning. Although current results are not as promising as the state-of-the-art method. We hope our experiments are inspirational and insightful for future researchers and scholars.

ACKNOWLEDGMENT

The authors would like to thank Dr. Hanqing Chao for his support on experimental result of some methods.

REFERENCES

[1] A. Sepas-Moghaddam and A. Etemad, "Deep gait recognition: A survey," *arXiv preprint arXiv:2102.09546*, 2021.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[4] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 4. IEEE, 2006, pp. 441–444.

[5] A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, "How to train your vit? data, augmentation, and regularization in vision transformers," *arXiv preprint arXiv:2106.10270*, 2021.

[6] Z. Zhang, L. Tran, X. Yin, Y. Atoum, X. Liu, J. Wan, and N. Wang, "Gait recognition via disentangled representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4710–4719.

[7] H. Chao, Y. He, J. Zhang, and J. Feng, "Gaitset: Regarding gait as a set for cross-view gait recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 8126–8133.

[8] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J. Chi, Y. Huang, Q. Li, and Z. He, "Gaitpart: Temporal part-based model for gait recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 14 225–14 233.

[9] S. Hou, C. Cao, X. Liu, and Y. Huang, "Gait lateral network: Learning discriminative and compact representations for gait recognition," in *European Conference on Computer Vision*. Springer, 2020, pp. 382–398.

[10] B. Lin, S. Zhang, and F. Bao, "Gait recognition with multiple-temporal-scale 3d convolutional neural network," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3054–3062.

Xuhang Chen received the M.Sc. degree from University of Pennsylvania, Philadelphia, PA, USA, in 2019. He is currently pursuing the Ph.D. degree with the Department of Computer and Information Science, University of Macau, Macau. His current research interests include computer vision and multimedia.

Chi-Man Pun (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in software engineering from the University of Macau in 1995 and 1998 respectively, and the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong in 2002. He was the Head of the Department of Computer and Information Science from 2014 to 2019. He is currently a Professor of computer and information science and the In Charge of the Image Processing and Pattern Recognition Laboratory, Faculty of Science and Technology, University of Macau. He has investigated many external funded research projects as PI and has authored/coauthored more than 200 refereed papers in top-tier journals, such as IEEE TRANSACTIONS ON PATTERN ANALYS IS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, and IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING and conferences, such as ACM MM, ECCV, AAAI, ICDE, and VR. His research interests include image processing and pattern recognition; information security and forensic for multimedia; adversarial machine learning and AI security; and so on. He has also served as the General Chair/General Co-Chair/Program Chair for many international conferences, including the IEEE International Conference on Visual Communications and Image Processing (VCIP2020), and a PC member for many top conferences, such as AAAI, ICCV, CVPR, and so on.