A. Brömme, N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb,
A. Sequeira, M. Todisco, and A. Uhl (Eds.): BIOSIG 2022,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2022

Incorporation of Extra Pseudo Labels for CNN-based Gait Recognition

Daigo Muramatsu[1], Kousuke Moriwaki[2], Yoshiki Maruya[3], Noriko Takemura[4], Yasushi Yagi[5]

Abstract: CNN is a major model used for image-based recognition tasks, including gait recognition, and many CNN-based network structures and/or learning frameworks have been proposed. Among them, we focus on approaches that use multiple labels for learning, typified by multi-task learning. These approaches are sometimes used to improve the accuracy of the main task by incorporating extra labels associated with sub-tasks. The incorporated labels for learning are usually selected from real tasks heuristically; for example, gender and/or age labels are incorporated together with subject identity labels. We take a different approach and consider a virtual task as a sub-task, and incorporate pseudo output labels together with labels associated with the main task and/or real task. In this paper, we focus on a gait-based person recognition task as the main task, and we discuss the effectiveness of virtual tasks with different pseudo labels for construction of a CNN-based gait feature extractor.

Keywords: Gait Recognition, Attribute, Pseudo label, CNN.

1 Introduction

Gait recognition is a biometric person authentication method that utilizes features extracted from walking motion for recognition. There are two major approaches for gait recognition: model-based approaches [BJ01, YNC04, AN12], and model-free/appearance-based approaches [HB06, BXG09, LCL11]. In this paper, we focus on appearance-based approaches.

Appearance-based approaches utilize walking image sequences and extract image-based gait descriptors, such as gait energy image (GEI) [HB06], gait entropy image (GEnI) [BXG09], and gait flow image (GFI) [LCL11], from the image sequences. Recently, employing convolutional neural networks (CNNs) is the mainstream approach [Sh16, Wu17, Ta18]. Extracted image-based descriptors are used as inputs to the CNN, and several types of network structures and/or input/output architectures together with loss functions have been discussed. These approaches only focus on a single task, where one type of label is used for CNN training.

Another approach for CNN-based methods is to incorporate multiple labels, typified by multi-task learning, where multiple labels/information associated with different tasks are used for CNN training and construction. Several researchers in a number of fields have

[1] Seikei University, 3-3-1 Kichijoujikitamachi, Musashino, Tokyo, Japan, muramatsu@st.seikei.ac.jp
[2] NEC Corporation, 1753 Shimonumabe, Nakahara-ku, Kawasaki, Kanagawa, Japan, k.moriwaki@nec.com
[3] Seikei University, 3-3-1 Kichijoujikitamachi, Musashino, Tokyo, Japan
[4] Kyushu Institute of Technology, 680-4 Kawazu, Iizuka-shi, Fukuoka, JAPAN, takemura@ai.kyutech.ac.jp
[5] Osaka University, 8-1 Mihogaoka, Ibaraki, Osaka, Japan, yagi@sanken.osaka-u.ac.jp

reported that the accuracy of the main task is improved by incorporating labels of different tasks together with those of the main task (e.g., [Zh14, Fu16, EY18, MYF19]). Multiple labels have also been incorporated in gait recognition; Marín et al. proposed multi-task learning for gait recognition by incorporating labels associated with age, gender and identity [Ma17]. They reported that this can contribute to faster training and can achieve better accuracy than that of a single task.

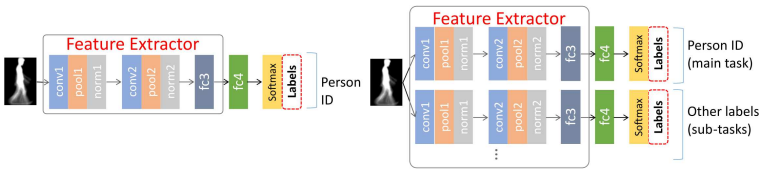

Fig. 1: (left) Network structure of GEINet. (right) Network structure of extended GEINet.

Incorporation of multiple labels can be effective for gait recognition, but the limitations of this method are that it necessitates extra labels for the target data, and appropriate sub-tasks are unknown in advance. In the case where the target data is annotated with several types of labels, this method is applicable, but it is unclear if available annotated label are useful for the target main task. And in the case where no extra annotated label is available, this method is not applicable; in order to apply it, we would need to annotate the data with labels even though we do not know what kind of label is appropriate for the main task.

In order to relax these limitations, in this paper, we consider incorporating pseudo labels for gait recognition. Our idea is that we generate virtual sub-tasks by annotating each item of data with a pseudo label, and use the virtual task to improve the accuracy of the main task. We focus on a CNN-based gait feature extractor named GEINet [Sh16] shown in Fig.1 (left), and extend GEINet using additional labels, as shown in Fig.1 (right). By this extension, we show the potential of pseudo labels for gait recognition [5]. Unlike the majority of image-based recognition tasks, output class labels of a CNN are not used in person recognition tasks, because available class labels for training are different from those for testing.

The contributions of this paper are summarized in the following three points:

- We propose an extended gait feature extractor incorporating extra labels. With real labels, we can extract features by considering additional labels, and can improve the gait recognition accuracy.
- We propose incorporating pseudo labels in the feature extractor and show that, by doing so, we can achieve an accuracy improvement comparable to that achieved with real labels associated with age and gender. This result is interesting because a pseudo label has no semantic meaning, but it can contribute to accuracy improvement if the label is annotated based on some constraint. Moreover, this approach has the significant advantage that annotation of additional labels for sub-tasks is unnecessary. This advantage is extremely promising, because we can apply this framework to all CNN-based feature extractors without having to prepare costly additional annotations.

[5] In this paper, we use GEINet as a backbone model, We can extend any type of backbone models using the same way.

- We can further improve the gait recognition accuracy by incorporating pseudo labels in addition to real labels, such as age and gender, or other pseudo labels. This result implies that there is a possibility of generating an optimal virtual sub-task by combining multiple virtual tasks.

2 Incorporation of extra label for recognition

Feature extracted from network

We focus on GEINet [Sh16] (shown in Fig. 1 (legt)) as a gait feature extractor. A GEI is input to GEINet, and the output of the full connection layer (fc3 in Fig. 1 (left)) is used as a gait feature vector. Based on this GEINet, extended GEINet (exGEINet) is generated by adding an extra subnetwork in parallel with the original network, as shown in Fig. 1 (right). The input to the added network is the same GEI as that input to GEINet, and output labels associated with the added network are set depending on the sub-tasks.

Let $F^{\mathrm{fc3}}(x;\mathbf{\Omega})$ be a feature vector (column vector) extracted from exGEINet with parameter set $\mathbf{\Omega}$ associated with an input GEI x. Let $f_n^{\mathrm{fc3}}(\cdot;\omega_i) \in \mathbb{R}^{K_n}$ be the output column vector from fc3 of the n-th subnetwork with parameter ω_n, which is associated with he n-th task, where K_n is the number of components in fc3 of the n-th subnetwork. When we consider N tasks including the main task (the main task corresponds to the 1st subnetwork), the feature extracted from input x is described by

$$F^{\mathrm{fc3}}(x;\mathbf{\Omega}) = \left[f_1^{\mathrm{fc3}}(x;\omega_1)^{\mathrm{T}}, f_2^{\mathrm{fc3}}(x;\omega_2)^{\mathrm{T}}, ..., f_N^{\mathrm{fc3}}(x;\omega_N)^{\mathrm{T}} \right], \tag{1}$$

where superscript "T" denotes the transposition operator. The dimension of the extracted feature is $\sum_n K_n$.

Extra label for sub-tasks

As is clear from (1), the extracted feature is heavily dependent on the employed sub-tasks. Therefore, selection of the sub-tasks is a crucial issue for exGEINet. A simple solution to this issue is to select related real tasks heuristically. For example, gender classification and/or age-group classification can be candidates for gait recognition, as reported in [Ma17] if these labels are available. This solution is satisfactory, and as reported in Section 3, each of these tasks can improve the accuracy of the main task. This shows the efficiency of exGEINet. At the same time, a question arises: what kind of sub-tasks are useful for improving the accuracy of the main task. In order to find a clue, we consider not only real tasks, but also virtual tasks. Labels of the virtual tasks have no semantic meaning; each label is assigned following some rules.

In order to realize the virtual task, we first set the number of classes N_n associated with the n-th task, and then automatically annotate each item of data with a class label under the constraint that the data originating from the same subject must have the same label; in other words, we annotate each subject with a random label. Here different subjects can have the same label for a task, but one subject does not have a different label for a task. We refer to this constraint as the *subject constraint* in this paper. This constraint is important for the person recognition task.

Training

Let $\mathcal{D}_n$ be the training dataset of the n-th subnetwork that is composed of M pairs of data x_m and label l_m^n associated with the n-th task. In the case of a real task, l_m^n is an annotated one-hot vector for data x_m, and in the virtual task, l_m^n is also a one-hot vector in which the hot label is assigned randomly under the subject constraint.

In order to set the parameter ω_n of the subnetwork for the n-th task, including the main task, we optimize the following loss function, $L_n(\omega_n; \mathcal{D}_n)$, independently:

$$L_n(\omega_n; \mathcal{D}_n) = -\sum_{m=1}^{M}\sum_{i=1}^{N_n} \delta_{(l_m^n)_i,1} \log f_n^i(x_m; \omega_n). \tag{2}$$

Here, $f_n(x; \omega_n)$ is an output vector of the softmax function associated with the n-th subnetwork against input x, and $f_n^i(x; \omega_n)$ and $(l_m^n)_i$ are the i-th component of the output vector and the i-th component of the one-hot vector (l_m^n), respectively. δ is the Kronecker delta function with two variables. One of the variables is set to 1 in (2), and δ becomes 1 only when $(l_m^n)_i = 1$. We denote the optimized parameter set of the n-th task as ω_n^*. It should be noted that training of each subnetwork is completely independent, and therefore, this approach is different from soft parameter sharing like that in [Du15, YH17].

Matching

Let x_g and x_p be the gallery GEI and the probe GEI, respectively. In the person recognition task, we extract features using exGEINet and make a decision based on a dissimilarity score calculated from two input features. The dissimilarity score between two GEIs x_g and x_p is computed by

$$d(x_p, x_g; \Omega^*) = \left\| F^{\mathrm{fc3}}(x_p; \Omega^*) - F^{\mathrm{fc3}}(x_p; \Omega^*) \right\|_2, \tag{3}$$

where $\Omega^* = \left(\omega_1^*, \omega_2^*, ..., \omega_N^*\right)$ and $\|\cdot\|_2$ means L2 norm. The dissimilarity score between the input probe and each stored gallery is calculated, and calculated dissimilarity scores are used for person identification.

3 Experiment

Dataset

A subset of the OU-ISIR gait database comprising the large population dataset (OU-LP) [Iw12] was used for evaluation. A subset of OU-LP was composed of silhouette image sequences from 1912 subjects with four different azimuth viewing angles: 55, 65, 75 and 85 deg. At each viewing angle and each subject, two silhouette image sequences were available, and age and gender labels of each subject were also available.

For accuracy evaluation, we divided subjects in the dataset into two groups: training subjects and test subjects. GEIs associated with the training subjects were used to train parameters of exGEINet, and GEIs associated with the test subjects were used for identification accuracy evaluation. It should be noted that there was no overlap between the training subjects and the test subjects.

Implementation

The layer configuration of each subnetwork is summarized in table 1. For the number of components of fc3, K_n, associated with the n-th task, we set $K_n = 1024$ for all of the tasks. exGEINet is composed of N subnetworks, and we varied the number of subnetworks from 1 (this corresponds to GEINet) to 4 (one main task + three sub-tasks). The size of the input GEI was set to 88×128 pixels. GEINet and exGEINet were trained and tested using Caffe [Ji14].

Tab. 1: Layer configuration of subnetworks for exGEINet.

Layer	#Kernels	Size/stride	Activation	Pooling
conv1	18	7×7×1/1	ReLU	-
pool1	-	2×2/2	-	Max pooling
conv2	45	5×5×18/1	ReLU -	
pool2	-	3×3/2	-	Max pooling

Sub-tasks

We considered the following four real tasks and three virtual tasks for the evaluation. For the real tasks, the following labels were used for construction of the subnetworks.

- Gender [2 classes]: Male (M), and Female (F)
- Life-stage (Ls) [6 classes]): (0-5), (6-14), (15-29), (30-44), (45-64), (over 65)
- Age group (Ag) [9 classes]: (0-9), (10-19), (20-29), (30-39), (40-49), (50-59), (60-69), (70-79), (over 80)
- Gender×Age group (GAg) [17 classes][6]

For the virtual tasks, we set the number of classes to 6, 9 and 18 so that the number of classes associated with the virtual tasks corresponded to those of real tasks. We refer to these tasks as Rd6, Rd9, and Rd18, respectively. It should be noted that the labels for the virtual tasks were assigned randomly under the subject constraint. There was no common attribute among the data with the same class label.

Experimental results

We set the viewing angle of the gallery to 85 deg and evaluated cross-view identification accuracy by changing the viewing angle of the probe to 55, 65, 75 and 85 deg. For evaluation, we measured rank-1 identification rates and analysed the impact of each sub-task on the main task.

The rank-1 results are summarized in Table 2, and Fig. 2, and 3. From these results, we can observe the following:

1. When we considered a real task GAg, rank-1 improved in many cross-view settings. In particular, the accuracy on similar view settings was improved, as shown in Fig. 2 (left). These results show that incorporation of extra labels associated with age and gender could contribute to improved identification accuracy.

2. A virtual task with pseudo extra labels could contribute to improved accuracy in the majority of cross-view settings. In particular, Rd18 achieved much better accu-

[6] Originally this must be 18 classes, but data associated with "female over 80" was not included in the training data, and we considered 17 classes by removing this class.

Tab. 2: Rank-1 identification rates [%] when the viewing angle of the gallery was 85 deg.

	Viewing angle of probe [deg]			
Tasks (#class of sub-task)	55	65	75	85
ID only	79.9	90.6	93.8	94.7
ID+gender (2)	78.2	91.5	94.6	95.8
ID+Ls (6)	79.1	91.9	94.6	95.8
ID+Ag (9)	80.3	91.9	95.0	95.9
ID+GAg (17)	80.8	92.0	95.0	96.2
ID+Rd6 (6)	80.2	92.1	94.7	96.0
ID+Rd9 (9)	80.8	90.2	95.1	96.2
ID+Rd18 (18)	82.6	93.8	95.8	96.5
ID+Gendar+Rd18 (2+18)	82.8	93.8	95.8	96.6
ID+Ag+Rd18 (9+18)	83.8	94.0	96.3	96.9
ID+GAg+Rd18 (2+9+18)	83.9	94.1	96.6	97.1
ID+Rd6+Rd18 (6+18)	84.1	94.0	96.1	96.8
ID+Rd9+Rd18 (9+18)	85.1	94.3	96.5	96.9
ID+Rd6+Rd9+Rd18 (6+9+18)	85.4	94.3	96.7	97.0

racy improvement. We observed the trend that virtual tasks with a larger number of classes tended to achieve better accuracy, as shown in Fig. 2 (right). An interesting observation is that virtual tasks with pseudo labels achieved accuracy comparable to or better than real tasks with age or gender, as shown in Fig. 2 (right). These results show that age or gender is not an optimal extra label for gait identification.

3. By combining a real sub-task and a virtual sub-task together with the main task, the identification accuracy was further improved. For example, the identification accuracy was improved from 79.9 % to 83.9 % in a setting where the viewing angle of the probe was 55 deg and the viewing angle of the gallery was 85 deg by a combination of ID, GAg, and Rd18. Moreover, combining virtual tasks together with the main task achieved the best or the second-best accuracy. This result shows the possibility that we can generate a more appropriate virtual task by combining multiple pseudo labels.

4 Conclusion

In this paper, we focus on a CNN-based gait feature extractor for person identification, and we propose an extended CNN-based feature extractor called exGEINet by incorporating extra labels. For the extra labels, we consider pseudo labels together with real labels such as gender and age, and we evaluate the efficiency of exGEINet. Experimental results using the OULP gait dataset showed that not only the labels associated with real tasks but also pseudo labels can contribute to improved accuracy of gait recognition. Moreover, incorporating multiple extra labels can achieve a further improvement.

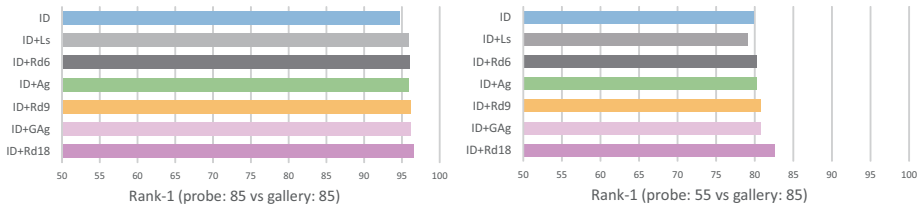

Fig. 2: Impact of incorporating one-type of extra label for gait identification: (left) when the viewing angle of the probe was 85 deg and the viewing angle of the gallery was 85 deg. (right) when the viewing angle of the probe was 55 deg and the viewing angle of the gallery was 85 deg.

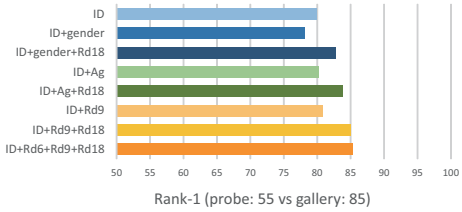

Fig. 3: Impact of incorporating multiple extra labels for gait identification when the viewing angle of the probe was 55 deg and the viewing angle of the gallery was 85 deg.

In the case where we use labels associated with real tasks, additional manual annotation or additional information is necessary, but in the case where we use pseudo labels, we can produce multiple types of pseudo labels without manual annotation. This is an advantage of incorporating pseudo labels, because we can generate different types of virtual tasks by combining multiple pseudo labels.

This is the first step of our research; we only consider randomly selected pseudo labels, and simply concatenate multiple subnetworks for the feature extractor. The number of fc3 associated with each subnetwork is fixed, and the number of classes in each task is set heuristically. We plan to tackle these issues in future work.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 17H02000 and 20H04188.

References

[AN12] Ariyanto, G.; Nixon, M.S.: Marionette mass-spring model for 3D gait biometrics. In: Proc. of the 5th IAPR Int. Conf. on Biometrics. pp. 354–359, March 2012.

[BJ01] Bobick, A.F.; Johnson, A.Y.: Gait Recognition using Static Activity-specific Parameters. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition. volume 1, pp. 423–430, 2001.

[BXG09] Bashira, K.; Xiang, T.; Gong, S.: Gait recognition using gait entropy image. In: Proc. of the 3rd Int. Conf. on Imaging for Crime Detection and Prevention. pp. 1–6, Dec. 2009.

[Du15] Duong, Long; Cohn, Trevor; Bird, Steven; Cook, Paul: Low Resource Dependency Parsing: Cross-lingual Parameter Sharing in a Neural Network Parser. In: Proceedings of the 53rd

Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Association for Computational Linguistics, pp. 845–850, 2015.

[EY18] Ege, Takumi; Yanai, Keiji: Multi-task Learning of Dish Detection and Calorie Estimation. In: Proceedings of the Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management. CEA/MADiMa '18, ACM, New York, NY, USA, pp. 53–58, 2018.

[Fu16] Fukui, Hiroshi; Yamashita, Takayoshi; Kato, Yuu; Matsui, Ryo; Ogata, Tetsuya; Yamauchi, Yuji; Fujiyoshi, Hironobu: Multiple Facial Attributes Estimation Based on Weighted Heterogeneous Learning. In: ACCV Workshops. 2016.

[HB06] Han, J.; Bhanu, B.: Individual Recognition Using Gait Energy Image. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(2):316–322, 2006.

[Iw12] Iwama, H.; Okumura, M.; Makihara, Y.; Yagi, Y.: The OU-ISIR Gait Database Comprising the Large Population Dataset and Performance Evaluation of Gait Recognition. IEEE Transactions on Information Forensics and Security, 7(5):1511–1521, Oct 2012.

[Ji14] Jia, Yangqing; Shelhamer, Evan; Donahue, Jeff; Karayev, Sergey; Long, Jonathan; Girshick, Ross; Guadarrama, Sergio; Darrell, Trevor: Caffe: Convolutional Architecture for Fast Feature Embedding. In: Proceedings of the 22Nd ACM International Conference on Multimedia. MM '14, ACM, New York, NY, USA, pp. 675–678, 2014.

[LCL11] Lam, T. H. W.; Cheung, K. H.; Liu, J. N. K.: Gait flow image: A silhouette-based gait representation for human identification. Pattern Recognition, 44:973–987, April 2011.

[Ma17] Marin-Jimenez, M. J.; Castro, F. M.; Guil, N.; de la Torre, F.; Medina-Carnicer, R.: Deep multi-task learning for gait-based biometrics. In: 2017 IEEE International Conference on Image Processing (ICIP). pp. 106–110, Sep. 2017.

[MYF19] Matsui, Ryo; Yamashita, Takayoshi; Fujiyoshi, Hironobu: Simultaneous Estimation of Facial Landmark and Attributes with Separation Multi-task Networks. pp. 265–272, 01 2019.

[Sh16] Shiraga, K.; Makihara, Y.; Muramatsu, D.; Echigo, T.; Yagi, Y.: GEINet: View-invariant gait recognition using a convolutional neural network. In: 2016 International Conference on Biometrics (ICB). pp. 1–8, June 2016.

[Ta18] Takemura, N.; Makihara, Y.; Muramatsu, D.; Echigo, T.; Yagi, Y.: On Input/Output Architectures for Convolutional Neural Network-Based Cross-View Gait Recognition. IEEE Transactions on Circuits and Systems for Video Technology, pp. 1–1, 2018.

[Wu17] Wu, Z.; Huang, Y.; Wang, L.; Wang, X.; Tan, T.: 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, Feb 2017.

[YH17] Yang, Yongxin; Hospedales, Timothy: Trace Norm Regularised Deep Multi-Task Learning. In: 5th International Conference on Learning Representations Workshop. 2017.

[YNC04] Yam, C.; Nixon, M.S.; Carter, J.N.: Automated Person Recognition by Walking and Running via Model-based Approaches. Pattern Recognition, 37(5):1057–1072, 2004.

[Zh14] Zhang, Zhanpeng; Luo, Ping; Loy, Chen Change; Tang, Xiaoou: Facial Landmark Detection by Deep Multi-task Learning. In (Fleet, David; Pajdla, Tomas; Schiele, Bernt; Tuytelaars, Tinne, eds): Computer Vision – ECCV 2014. Springer International Publishing, Cham, pp. 94–108, 2014.