CHD: Consecutive Horizontal Dropout for Human Gait Feature Extraction

Chengtao Cai
College of Automation
Harbin Engineering University
Harbin, China, 150001
86045182589656
cai.chengtao@hrbeu.edu.cn

Yueyuan Zhou
College of Automation
Harbin Engineering University
Harbin, China, 150001
8615104555555
yueyuan_zhou@hrbeu.edu.cn

Yanning Wang
Faculty of Information Technology
Beijing University of Technology
Beijing, China, 100124
8619801355507
wangyym@emails.bjut.edu.cn

ABSTRACT
Despite gait recognition and person re-identification researches have made a lot of progress, the accuracy of identification is not high enough in some specific situations, for example, people carrying bags or changing coats. In order to alleviate above situations, we propose a simple but effective Consecutive Horizontal Dropout (CHD) method apply on human feature extraction in deep learning network to avoid overfitting. Within the CHD, we intensify the robust of deep learning network for cross-view gait recognition and person re-identification. The experiments illustrate that the rank-1 accuracy on cross-view gait recognition task has been increased about 10% from 68.9% to 78.2%. However, 3D template based method reduces the occupation of computing resources, making it possible to implement real-time gait recognition. The most popular template is gait energy image (GEI) which Generating a template image of silhouette images. The most popular template is gait energy image (GEI) [3]. Template-based method reduces the occupation of computing resources, making it possible to implement recognition based on gait information under limited hardware technologies. After template-based method, with the development of hardware technology, 3D convolutional neural network (CNN) has used for gait recognition [4, 5]. The input of 3D-CNN is 4-dimensional data of ordered gait sequence within temporal information, which retain more valuable information than template-based method and improve the accuracy of recognition. However, 3D-CNN need much more capability of calculation and memory than the template-based method so that 3D-CNN method is high-cost both on time and economy. Based on the above two methods, the third method loses less information while avoiding 3D-CNN, which has a trend of being widely used. The successful paradigm of the third method is Gaitset [6] which achieves the highest recognition accuracy. Therefore, we study based on the Gaitset.

However, with the three methods above can obtain a good accuracy on normal situations but worse accuracy on those special environments, for example people wearing coat, carrying bags, walking too fast, etc. In order to improve the robustness, many algorithms have been proposed. For example, HPM[7] and PS[8] deal with the missing body cases while Hierarchical Gaussian Descriptor[9] focus on color and textural. At the same time, we propose Horizontal Dropout (HD) committing to solving the poor robustness of gait recognition problem as well.

2. RELATED WORK
2.1 Deep Learning for Person Re-ID
The Re-ID can be considered as a matching or multi-classification mission in the computer vision[10]. The data always format as \( S = \{ S_1, S_2, ..., S_n \} \), where \( n \) is the number of sequences and \( s \) represent sequence number.

The sequence concluding \( n(s) \) human images are represented as:
\[
S_5 = \{ f_1^1, f_1^2, ..., f^n_1(s), f_2^1, f_2^2, ..., f^n_2(s), ..., f^n_s(s) \}
\]

With the above expression, Re-ID is divided into two types of model.

2.1.1 Identification model
Identification model regards Re-ID mission as classification task. The output of model is the most possible person ID of the input data. Due to the fact that identification models make full use of the training data, more and more algorithms based on identification model achieve great accuracy, for example, Y Lin et al.[11] work, D Li et al.[12] work and Spindle[13].

2.1.2 Verification model
Verification model usually need two or more input data, then compare the distance between probe and each input image to decide whether they are same person or not. PRDC[14] is the typical verification model which purpose is to make same label distance less than disparate one by reducing the distance of same ID and aggrandizing the distance of different ID label. Inspired by PRDC, D Cheng et al.[15], A Hermans et al.[16], W Chen et al.[17] addressed improved algorithm and achieved better performance.

2.2 Gait Recognition
Gait recognition can base on various sensor devices such as floor sensor system[18], depth sensors[19], acceleration sensors[20-23], and even WIFI signall[24]. We only cover vision-based gait
recognition which is also the most popular one. The gait recognition task always pretreats human images to silhouettes, that is different from person re-identification. Inchonate gait recognition only considered the 90° case where the person walked parallel to the camera[25, 26]. Recently, cross view gait recognition has been studied more and had some achievements [6, 27, 28].

2.3 Dropout

Dropout was designed by N Srivastave et al.[29] to avoid overfitting in the deep neural network training by setting neurons to zero randomly. In the next few years, many applications using dropout to improve various neural network successfully, such as N Srivastave et al.[30], GE Dahl et al.[31], Y Gai et al.[32], V Pham et al.[33] and H Wu et al.[34]. Furthermore, some people are working on improving dropout. Z Li et al.[35] proposed an efficient adaptive dropout in order to deal with the issue of evolving distribution of neurons in deep learning.

3. PROPOSED APPROACH

In this section we will describe the structure and implementation of our approach Consecutive Horizontal Dropout (CHD). However, at first, we designed two HD structures (see Figure 1) which are Consecutive HD (CHD) and Sporadic HD (SHD).

![Figure 1. The two structures of HD where CHD performs better than SHD even though SHD works as well.](image)

Actually, the SHD is more like traditional 3D dropout. But, the experiments in Section 4 indicate that the CHD is much better than SHD so that we recommend CHD. In theory, features of each row are not irrelevant. Before we do HD on feature, each row is a vector that represents a large view of spatial information on temporal dimension. If we split features to too many slides then the network cannot recognize the effective areas.

3.1 The structure of HD

The input of HD can be diverse human images such like RGB images and silhouettes. The arrow with CNN in Figure 1 represents convolutional neural network to extract features from input image as a backbone network. The features output from CNN will be used to calculate maximum and mean value on width dimension separately that reveal as mint cube and pink cube in Figure 1. After all these operations, we divide HD into two structures according to drop modality. The first is CHD as seen from (a) in Figure 1 and (b) in Figure 1 shows the second algorithm of SHD. CHD randomly zero out entire channels of a certain height while the SHD zero out entire channels on randomly rows that may not continuous. Black cube in Figure 1 express zeros. The number of row drops is a parameter to be decided in specific application. For a clearer description, the number of row drops will be represented as drop-number in the following sections. Finally, we add the maximum matrix and mean matrix to generate new features shown as the feldspar cube in Figure 1.

3.2 The Implementation of CHD

3.2.1 Definitions

In practical applications, multiple images are usually used as a batch for training, so that we express the input of a batch with n image(s) as $\text{Img}(n)$. Analogously, the output of CNN is denoted as a matrix $\text{In}(n, c, h, w)$, where $(n, c, h, w)$ denotes the shape of matrix $\text{In}$. Then we obtain two matrices $\text{Max}(n, c, h, 1)$ and $\text{Mean}(n, c, h, 1)$ and two dropped matrices $\text{Max}'(n, c, h, 1)$ and $\text{Mean}'(n, c, h, 1)$ according to the HD structure. Finally, the sum of $\text{Max}'$ and $\text{Mean}'$ is the matrix $\text{Sum}(n, c, h, 1)$. In addition, we always regard $(n, c, h, 1)$ and $(n, c, h)$ as the same shape in this paper.

3.2.2 Details

To verify the effectiveness of HD, we choose GaitSet without HPM to do the experiment, however all existing CNN backbone could be used, for example, Resnet[36], VGG[37], DenseNet[38] and etc. Please note that whatever the backbone is, the output should be 4 dimensional features including the batch dimension. When we obtain the output features from CNN backbone network, we have three options to implement exactly same HD (see Figure 2). In this paper we choose option (c) in Figure 2.

First, we obtain original features $\text{In}$ from modified GaitSet shown in Figure 3 with parameters in Table 1.

| Table 1. Parameters of training. |
|----------------------------------|
| Learning-rate | 0.001 |
| Batch-size | 11x16 |
| GPUs | 2 Tesla P100 |
| margin | 0.2 |
| Frame-number | 30 |
| Date-shuffle | False |
| Iterations | Depends |

Second, $\text{Max}$ and $\text{Mean}$ matrix are calculated with Formula 1 and Formula 2, where 3 denotes the third dimension count from zero.

\[
\text{Max}(n, c, h) = \max(\text{In}(n, c, h, w), 3)
\]

Formula 1.

\[
\text{Mean}(n, c, h) = \text{mean}(\text{In}(n, c, h, w), 3)
\]

Formula 2.

Third, we add $\text{Max}$ and $\text{Mean}$ matrix following the Formula 3.

\[
\text{Sum}(n, c, h) = \text{Max}(n, c, h) + \text{Mean}(n, c, h)
\]

Formula 3.

Then we generate a number as start number randomly. Finally, we zero the entire horizontal planes from row of start number to sum of start number and drop-number.
4. Experiments

4.1 Dataset

The dataset we use for training and testing our approach is CASIA-B [39] dataset which conclude 124 people over 15 thousand videos. Each person has 3 walking conditions with 10 captures which are “bg-01”, “bg-02”, “cl-01”, “cl-02”, “nm-01”, “nm-02”, “nm-03”, “nm-04”, “nm-05” and “nm-06”. LT, MT, and ST are 3 ways to divided training set and testing set. LT means large-sample training concluding first 74 people videos, so that the rest of 50 people’s videos are used to test. Analogously, MT represent middle-sample training with 62 people and ST denotes small-sample training with 24 people in training set.

All the experiments we do using LT. For the test set, we use “nm-01”, “nm-02”, “nm-03”, “nm-04” as gallery then employ “bg-01”, “bg-02”, “cl-01”, “cl-02” “nm-04”, “nm-05” and “nm-06” as probe.

4.2 Backbone

Essentially, the choice of backbone seriously affects the accuracy of experiment results. In order to prove the effect of our method, all experiment choose the newest and best network so far which is GaitSet. As a backbone network, we delete the HPP and concatenate the MGP features and main pipeline features together on the third dimension. Thus the output of the backbone is a matrix $In(n, c, h, w)$ where $h=32$ so that in our experiments drop-number is no more than 32.

Table 2. The rank-1 identification rates [%] of cross-view gait recognition with CHD

| Drop-number | NM   | BG   | CL   | Average |
|-------------|------|------|------|---------|
| 0           | 94.116 | 86.577 | 68.0  | 82.898  |
| 1           | 94.917 | 86.766 | 72.215 | 84.633  |
| 2           | 95.289 | 89.135 | 75.777 | 86.734  |
| 3           | 95.132 | 88.771 | 73.347 | 85.750  |
| 4           | 95.041 | 89.168 | 73.099 | 85.769  |
| 5           | 94.752 | 87.643 | 73.868 | 85.421  |
| 6           | 95.421 | 87.569 | 72.38  | 85.123  |
| 7           | 94.967 | 87.827 | 75.322 | 86.039  |
| 10          | 95.165 | 87.651 | 77.455 | 86.757  |
| 13          | 95.017 | 86.461 | 76.364 | 85.947  |
| 16          | 94.479 | 88.281 | 78.201 | **86.987** |
| 27          | 93.868 | 86.595 | 75.471 | 85.311  |
| 30          | 90.876 | 79.323 | 64.446 | 78.215  |
| 31          | 91.694 | 81.468 | 64.587 | 79.250  |

4.3 Application on Cross-View Gait Recognition

4.3.1 Consecutive HD

Table 2 shows the partial rank-1 accuracy results of consecutive HD applying in the gait recognition task with different drop-number. Comparing with first line without using HD, the accuracy of following lines except the last two have been improved in all three conditions (NM, BG, CL). When the drop-number is equal to 16, the accuracy in CL condition increased by over 10%. Besides, the accuracy of BG condition has been improved a lot and there is still a litter increase of NM condition as well. As can
be seen from the last column, while the CL accuracy is increasing, the average accuracy is also increasing.

We aim to enhance the robustness of the CNN network to gain higher accuracy of the CL condition, so that we are inclined to choose 16. However, if improving the accuracy of NM or BG condition is the purpose, 2 or 4 drop-number should be chose according to Table 2.

4.3.2 Sporadic HD

Table 3 lists typical experiments results. The second to sixth lines reveal that the SHD increase the accuracy when drop-number is less than 30. Among them, when the drop-number is equal to 1, the accuracy is the best in all three conditions (NM, BG, CL).

| Drop-number | NM   | BG   | CL   | Average |
|-------------|------|------|------|---------|
| 0           | 94.116 | 86.577 | 68.0  | 82.898  |
| 1           | **95.909** | **89.895** | **69.975** | **85.260** |
| 2           | 95.562 | 89.505 | 69.512 | 84.860  |
| 3           | 95.248 | 88.587 | 69.529 | 84.455  |
| 10          | 95.488 | 89.549 | 67.884 | 84.307  |
| 15          | 95.612 | 89.227 | 67.512 | 84.117  |
| 30          | 94.124 | 87.522 | 64.529 | 82.058  |

However, comparing with Table 2, the disadvantage of SHD is obvious. As the drop-number increases, the accuracy remains almost unchanged and even decreases slightly in the later period.

4.4 Application on Person Re-identification

4.4.1 Consecutive HD

Table 4 shows the typical rank-1 accuracy results of consecutive HD applying in the person re-identification task with different drop-number. Obviously, using HD with proper drop-number increased accuracy about 8% in CL condition. In addition, the accuracies of all three conditions have been improved by consecutive HD, and it is the first time that 100% accuracy appear in the benchmark of the CASIA-B.

| Drop-number | NM   | BG   | CL   | Average |
|-------------|------|------|------|---------|
| 0           | 99.818 | 97.635 | 83.545 | 93.666  |
| 1           | **100.0** | **97.999** | **87.364** | **95.121** |
| 2           | **100.0** | **98.817** | **87.455** | **95.424** |
| 4           | 99.727 | 96.906 | 86.273 | 94.302  |
| 7           | **100.0** | **97.635** | **88.273** | **95.303** |
| 10          | **100.0** | **97.358** | **90.273** | **95.877** |
| 13          | **100.0** | **96.814** | **88.636** | **95.15** |
| 14          | **100.0** | **97.179** | **91.364** | **96.181** |
| 16          | **100.0** | **96.268** | **88.0** | **94.756** |
| 27          | **100.0** | **95.814** | **80.455** | **91.968** |
| 30          | **99.636** | **95.814** | **80.455** | **91.968** |

Table 4 shows the typical rank-1 accuracy results of consecutive HD applying in the person re-identification task with different drop-number. Obviously, using HD with proper drop-number increased accuracy about 8% in CL condition. In addition, the accuracies of all three conditions have been improved by consecutive HD, and it is the first time that 100% accuracy appear in the benchmark of the CASIA-B.

4.4.2 Sporadic HD

Table 5 demonstrate the partial rank-1 accuracy results of SHD applying in person re-identification task with different drop-number. The best result for NM condition is when drop-number is equal to 1, but BG is 2 and CL is 3. Considering that our purpose is to enhance the network recognition ability in CL condition and the average accuracy is highest when drop-number is 3, we regard 3 as the best drop-number.

| Drop-number | NM   | BG   | CL   | Average |
|-------------|------|------|------|---------|
| 0           | 99.818 | 97.635 | 83.545 | 93.666  |
| 1           | **100.0** | **98.361** | **84.545** | **94.302** |
| 2           | 99.818 | **99.272** | **82.818** | **93.969** |
| 3           | 99.909 | 98.454 | **85.636** | **94.666** |
| 10          | 99.818 | 98.545 | 81.545 | 93.303  |
| 15          | 99.818 | 98.544 | 81.0 | 93.121  |
| 30          | 99.545 | 96.544 | 76.182 | 90.757  |

Same as cross-view gait recognition task, in person identification task, CHD performs much better than SHD.

5. Conclusion

The two methods we proposed, whether CHD or SHD, are both obviously effective in theory and experiments to extract human features. However, due to the fact that the rows of features are relevant, SHD will break the connection of each rows when drop-number is too large, we consider that CHD is more effective to enhance the generalization ability of deep learning network. In the future, we look forward to more applications of CHD.

6. ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No. 61673129, 51674109).

7. REFERENCES

[1] Chellappa, Rama, Ashok Veeraraghavan, and Narayanan Ramanathan. "Gait biometrics, overview." Encyclopedia of Biometrics (2015): 783-789.
[2] Wolf, Thomas, M. Babaei, and G. Rigoll. "Multi-view gait recognition using 3D convolutional neural networks." IEEE International Conference on Image Processing IEEE, 2016.
[3] Han, Ju, and Bir Bhanu. "Individual recognition using gait energy image." IEEE transactions on pattern analysis and machine intelligence 28.2 (2005): 316-322.
[4] Zhao, Guoying, et al. "3D gait recognition using multiple cameras." 7th International Conference on Automatic Face and Gesture Recognition (FG2006). IEEE, 2006.
[5] Wolf, Thomas, Mohammadreza Babaei, and Gerhard Rigoll. "Multi-view gait recognition using 3D convolutional neural networks." 2016 IEEE International Conference on Image Processing (ICIP). IEEE, 2016.
[6] Chao H, He Y, Zhang J, et al. GaitSet: Regarding Gait as a Set for Cross-View Gait Recognition[J]. 2018.
[7] Fu Y, Wei Y, Zhou Y, et al. Horizontal Pyramid Matching for Person Re-identification[J]. 2018.
[8] Cheng, Dong Seon, et al. "Custom pictorial structures for re-identification." Bmvc. Vol. 1. No. 2. 2011.
[9] Matsukawa, Tetsu, et al. "Hierarchical gaussian descriptor for person re-identification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
wearable sensors. Adv. Sci. Technology. Eng. System
investigation of human gait recognition database using
Savatier, X., & Abdrabbo, M. F. (2018). Experimental
gait recognition. In 2007 2nd IEEE Conference on Industrial
May). A wearable acceleration sensor system for gait
Conference on Imaging and Visualization.
Biomechanics and Biomedical Engineering and 3rd
Human Recognit
Advanced Technologies (AutoID’05) (pp. 171
In Fourth IEEE Workshop on Automatic Identification
Computer Vision and Pattern Recognition (pp. 403
Chen, W., Chen, X., Zhang, J., & Huang, K. (2017). Beyond
triplet loss: a deep quadruplet network for person
Identification. In Proceedings of the IEEE Conference on
Computer Vision and Pattern Recognition (pp. 403-412).
Middleton, L., Buss, A. A., Bazin, A., & Nixon, M. S.
(2005, October). A floor sensor system for gait recognition.
In Fourth IEEE Workshop on Automatic Identification
Advanced Technologies (AutoID’05) (pp. 171-176). IEEE.
Nunes, J. F., Moreira, P. M., & Tavares, J. M. R. (2018).
Human Recognition and Classification Based on Gait
Analysis Using Depth Sensors. In CMBBE2018-15th
International Symposium Computer Methods in
Biomechanics and Biomedical Engineering and 3rd
Conference on Imaging and Visualization.
Rong, L., Jianzhong, Z., Ming, L., & Xiangfeng, H. (2007,
May). A wearable acceleration sensor system for gait
recognition. In 2007 2nd IEEE Conference on Industrial
Electronics and Applications (pp. 2654-2659). IEEE.
Gafurov, D., Helkala, K., & Sondrol, T. (2006). Biometric
Gait Authentication Using Accelerometer Sensor. JCP, 1(7),
51-59.
Gafurov, D., Snekkenes, E., & Bours, P. (2007, June). Gait
authentication and identification using wearable
accelerometer sensor. In 2007 IEEE workshop on automatic
identification advanced technologies (pp. 220-225). IEEE.
Said, S., Al-Kork, S., Nair, V., Gowthami, I., Beyrouthy, T.,
Savatier, X., & Abdrabbo, M. F. (2018). Experimental
investigation of human gait recognition database using
wearable sensors. Adv. Sci. Technology. Eng. System
Journal, 3(4), 201-210.
Nipu, M. N. A., Talukder, S., Islam, M. S., & Chakrabarty,
A. (2018, June). Human Identification Using WIFI Signal. In
2018 Joint 7th International Conference on Informatics,
Electronics & Vision (ICIEV) and 2018 2nd International
Conference on Imaging, Vision & Pattern Recognition
(icIVPR) (pp. 300-304). IEEE.
Lee, L., & Grimson, W. E. L. (2002, May). Gait analysis for
recognition and classification. In Proceedings of Fifth IEEE
International Conference on Automatic Face Gesture
Recognition (pp. 155-162). IEEE.
Wang, L., Tan, T., Ning, H., & Hu, W. (2003). Silhouette
analysis-based gait recognition for human identification.
IEEE transactions on pattern analysis and machine
intelligence, 25(12), 1505-1518.
Tong, S. B., Fu, Y. Z., & Ling, H. F. (2019). Cross-view
gait recognition based on a restrictive triplet network. Pattern
Recognition Letters, 125, 212-219.
Ben, X., Zhang, P., Lai, Z., Yan, R., Zhai, X., & Meng, W.
(2019). A general tensor representation framework for
cross-view gait identification. Pattern Recognition, 90, 87-98.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I.,
& Salakhutdinov, R. (2014). Dropout: a simple way to prevent
neural networks from overfitting. The journal of machine
learning research, 15(1), 1929-1958.
Srivastava, N. (2013). Improving neural networks with
dropout. University of Toronto, 182, 566.
Dahl, G. E., Sainath, T. N., & Hinton, G. E. (2013, May).
Improving deep neural networks for LVCSR using rectified
linear units and dropout. In 2013 IEEE international
conference on acoustics, speech and signal processing (pp.
8609-8613). IEEE.
Gal, Y., & Ghahramani, Z. (2016). A theoretically grounded
application of dropout in recurrent neural networks. In
Advances in neural information processing systems (pp.
1019-1027).
Pham, V., Bluche, T., Kermorvant, C., & Louradour, J.
(2014, September). Dropout improves recurrent neural
networks for handwriting recognition. In 2014 14th
International Conference on Frontiers in Handwriting
Recognition (pp. 285-290). IEEE.
Wu, H., & Gu, X. (2015). Towards dropout training for
convolutional neural networks. Neural Networks, 71, 1-10.
Li, Z., Gong, B., & Yang, T. (2016). Improved dropout for
shallow and deep learning. In Advances in Neural
Information Processing Systems (pp. 2523-2531).
He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual
learning for image recognition. In Proceedings of the IEEE
conference on computer vision and pattern recognition (pp.
770-778).
Simonyan, K., & Zisserman, A. (2014). Very deep
convolutional networks for large-scale image recognition.
arXiv preprint arXiv:1409.1556.
Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K.
Q. (2017). Densely connected convolutional networks. In
Proceedings of the IEEE conference on computer vision and
pattern recognition (pp. 4700-4708).
Yu, S., Tan, D., & Tan, T. (2006, August). 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, pp. 441-444). IEEE.