Pedestrian age recognition method based on gait deep learning

Tao Song¹*, Xin Yang¹, Yulin Wang², Yu Lei¹, Guanting Liu¹
¹Chongqing university of technology, Chongqing, 400054, China
²Chongqing vocational college of public transportation, Chongqing, 402247, China
*Corresponding author’s E-mail: tsong@cqut.edu.cn

Abstract. A pedestrian age recognition method based on gait information and dynamic weight convolutional networks is proposed to solve the problem that face age recognition needs a close positive perspective and frontal angle which limits its application scenarios. Firstly, the method takes gait energy image as input and the gait energy image is divided into head, upper body and lower body. Then, the global and local features of gait are extracted by global local convolution networks. Secondly, the multi-layer residual expansion networks is constructed to optimize the gait energy image and the average pooling is used to replace the fully convolutional networks for age classification. In order to solve the problem of sample imbalance in different age groups, the dynamic weight cross-entropy loss function is utilized as the network loss function to constrain the parameter updating, so as to avoid the recognition results inclining to the class with more samples. Experimental results on OULP age dataset demonstrate the effectiveness of the proposed approach.

1. Introduction
As an important attribute feature of pedestrians, age recognition has been widely used in AI era. At present, the mainstream method is age recognition based on face images [1–3], but this method has great limitations and is suitable for close-range active cooperation scenarios. However, in more occasions, the face image cannot be obtained from the front in close range or the face is covered, which brings challenges to age recognition. Extensive studies have shown that pedestrian gait contains potential biological information, including information such as gender and age. In the field of gait recognition, binary images are usually used to represent the human contour, and gait energy image (GEI) is used to characterize the pedestrian features. The human gait can be obtained from a long distance without the cooperation of the target, thus possessing extremely high concealment [4–5]. Han et al. [6] first introduced gait energy image, and applications based on gait features had been widely studied since then, such as access control, monitoring [7], medical treatment [8]. Osaka University in Japan constructed a whole-generation gait database, OULP-Age [9], and proved the feasibility of age estimation based on gait. Ince et al. [10] successfully distinguished children from adults using the simple ratio of head to body size. Nabila et al. [11] successfully distinguished the elderly from adults by using the curves obtained from the longitudinal and transverse projections of the binary human body contour image as their features. Deep learning also promotes the development of gait recognition [12–13]. Marin-Jiménez et al. [14] designed a multi-task convolutional neural networks (CNN), which regarded optical flow as a training set and realized the recognition of target identity, gender and age at the same time. However, only 305 individuals were included in the sample set, and most of these individuals were in their twenties, which was not universal.
In this paper, the gait database OULP-Age is taken as a sample, and the gait energy image and deep convolutional networks are used to identify the age. This database is the largest gait sample set at present. However, the number of samples of different ages and their imbalance will directly affect the recognition results of deep learning. To move beyond such limitations, a multi-classification cross-entropy loss function with dynamic weights is designed based on the multi-classification cross-entropy function. The specific loss function is used to train the residual expansion convolutional networks, and extensive experimental results verify that the proposed method effectively improves the recognition rate of gait age.

2. Related work

2.1. Age recognition methods based on hand-crafted feature

Early age recognition consists of two steps: feature extraction and classifier training. Researchers usually use manual design operators to effectively extract shallow visual features such as texture and contour, for example, Histogram of Oriented Gradient (HOG) features and Local Binary Pattern (LBP) features from facial images [15]. Cao et al. [16] used the Canny operator to obtain the edge features of the pedestrian image and combined with HOG features to obtain a more robust representation of the lighting conditions and locations. Deng et al. [17] trained SVM and CRF classifiers to perform attribute recognition based on the color and texture features of the pedestrian image.

2.2. Age recognition methods based on deep learning

Using the depth learning method can not only get the low-level features of the image in the early training, but also obtain the high-level semantic information to enhance the effectiveness of recognition along with the network training. Liu et al. [18] proposed a novel network guided by local information and used fused global-local features for attribute prediction. Zhao et al. [19] used CNN to extract the local information of the head, upper body and lower body respectively, grouped them according to the spatial distribution of attributes and used recurrent neural networks (RNN) to mine the connections of each attribute group to obtain the final recognition result.

In this paper, the gait information of pedestrians is combined with the residual expansion convolution networks to extract rich attribute features, and on this basis, proposed a dynamic weight cross-entropy loss function, which effectively solves the problem of poor network training effect caused by data imbalance.

3. Method

The purpose of convolution is to learn the representation of the input features. The convolutional layer consists of multiple convolution kernels. Mathematically, that eigenvalues \((i, j)\) in the \(k\)-th feature map of the \(l\)-th layer is calculated by the follow formula:

\[
Z_{i,j,k}^l = W_{i,k}^l X_{i,j}^l + b_k^l
\]

(1)

Where \(W_{i,k}^l\) and \(b_k^l\) are the \(k\)-th weight and offset items in the \(l\)-th layer, respectively, and \(X_{i,j}^l\) is the input values in the \(k\)-th feature map. The weight \(W_{i,k}^l\) for generating feature map \(Z_{i,j}^l\) is shared.

Let \(\theta\) denote all CNN parameters (weights and bias terms), and the best parameters for a specific task can be obtained by minimizing the loss function defined on the task. Suppose there are \(N\) input-output relationships, where \(x^{(n)}\) is the \(n\)-th input data, \(y^{(n)}\) is its corresponding target label, and \(o^{(n)}\) is the output of CNN. The loss function \(\ell\) can be calculated as follows:

\[
\ell = \frac{1}{N} \sum_{n=1}^N l(\theta; y^{(n)}, o^{(n)})
\]

(2)

The cross-entropy loss of one of the nodes is defined as follows:
\[
\ell_i = \sum_i -y_i \log(o_i)
\]  
(3)

Combined with Eqn.2, the final loss function becomes:
\[
\ell = \frac{1}{m} \sum_{j=1}^{l} \sum_{i} -y_{i,j} \log(o_{i,j})
\]  
(4)

Where \(m\) is the training batch size.

3.1. **Dynamic weight cross entropy loss function**

The deep neural networks is a complex nonlinear function in mathematics, which is assumed as follows:
\[
p = f(W, x)
\]  
(5)

Where \(p\) is the predicted value, \(W\) is the convolution weight of each layer, and \(x\) is the input. In order to get the correct predicted value, it is essential to minimize the loss function, which is mathematically expressed as follows:
\[
W^* = \arg \min_W \ell(W)
\]  
(6)

Compare the prediction with the label, calculate the loss, then calculate the partial derivative of the loss for each weight \(w_{i,j}\), and finally adjust the weight according to the partial derivative (gradient). In the \(l\)-th layer, suppose that \(Z^{l}_{j}\) is the original output of the \(k\)-th feature map, \(o^{l}_{i,j}\) is the output of the \(k\)-th feature map of the \(l\)-th layer after the activation function, and \(\ell\) is the loss function of the entire network. Then the weight iteration is as follows:
\[
W = W - \alpha \frac{\partial \ell}{\partial w}
\]  
(7)

where \(\alpha\) is learning rate.

In order to avoid this situation, this paper designs a dynamic weighted cross-entropy loss function, as shown in Eqn.8.
\[
\ell = \frac{1}{m} \sum_{j=1}^{l} \sum_{i} -y_{i,j} E_{j} \log(o_{i,j})
\]  
(8)

where \(E_{j}\) is the misidentification rate of class \(j\) in the current training batch. The specific calculation process is as follows:

Firstly, in each class, the misidentification rate \(E(e_1,e_2,...,e_i)\) in each training batch is counted. In order to keep the weight stable, calculate the cumulative average of each class:
\[
E = \frac{\sum_{k=1}^{K} E_{k}}{K}
\]  
(9)

Where \(K\) is the number of iterations of training. Assuming that the common cross-entropy loss function is expressed as \(\ell\), the dynamic cross-entropy loss function is expressed as \(E\ell\). Since \(E\) is a constant, the derivative of the loss function to \(l\)-th layer parameter \(w\) in back propagation is \(E \frac{\partial \ell}{\partial w}\), and the final parameter update calculation is as follows:
\[
W = W - \alpha E \frac{\partial \ell}{\partial w}
\]  
(10)

3.2. **Convolutional networks structure for gait feature age recognition**

**Gait energy image.** The background and foreground of the image are separated, and then the human body region is represented by a binary image, as shown in Figure 1.(a). Gait is periodic, and one period is taken as a set of valid data in gait recognition. By observing the gait silhouette sequence, it is found that the foreground pixels of the image reach the minimum and maximum respectively when the the
target’s feet are close together and when the feet are separated, which can be used as the basis for judging the gait cycle. Because the position of the human body area under a fixed visual angle is changing during walking, before generating GEI, the center of the binary silhouette image should be aligned to generate a centered silhouette, as shown in Figure 1.(b). Counting the sum of gray values of foreground pixels of each frame separately, and use this as gait energy map GEI, as shown in Figure 1.(c). Compared with binary contour sequence representation, GEI representation can save storage space and calculation time, and is less sensitive to contour noise in each frame. GEI is defined as formula:

\[ G(x, y) = \frac{1}{M} \sum_{i=1}^{M} B_i(x, y) \]  

(11)

Where \((x, y)\) is the centered coordinate, \(B_i\) is the gray value of the \(i\)-th frame image within one cycle, and \(M\) denotes the number of image frames within one gait cycle.

3.3. Convolutional networks structure

**Global-local convolutional networks.** In order to extract local and global features, Zhu H et al. [22] divided the gait energy image into head-gait energy image (H-GEI), chest-gait energy image (C-GEI) and leg-gait energy image (L-GEI) according to the scale of 0:22:70:128. Figure 2 shows the divided gait energy image. On this basis, the segmented energy images are convolved to obtain their own convolution feature maps, which are spliced into a complete convolution feature map according to the head-chest-leg sequence and stacked with the feature map obtained by convolution of the full-gait energy image. The convolutional layer is named as Global-Local Convolutional Layer (GLCL). The specific process is shown in Figure 3.

**Residual expansion networks.** Since He Kaiming and others put forward the residual networks, it has been widely used to solve the gradient disappearance problem of chain derivative and multiplication in neural network. Its basic principle is to construct an identity map, assume the function of fitting the neural network layer of deep convolution as \(H(\cdot)\), and split the function into:

\[ z' = H(d'z) = d'z + f(d'z) \]  

(14)

The residual block is divided into two parts: the direct mapping part and the residual part, in which \(d'z\) is the direct mapping part and \(f(d'z)\) is the residual part. Learning an identical mapping in the deeper network layer, the mapping after introducing the residual is more sensitive to the change of output and has a greater effect on weight adjustment. Adding residual units can be realized by layer-hopping connection method. Residual networks can solve the problem of network depth and accelerate the convergence speed of neural network training. However, the size of convolution template in convolutional neural networks is usually set according to experience. Drawing lessons from the idea of introducing Inception unit in GoogleNet, this paper expands the network width based on residual networks, uses templates of different sizes to convolute with the input, obtains multi-scale features, and stacks them in the channel direction. The residual expansion module designed in this paper is shown in Figure 4.
The input of the current layer is convoluted through a convolution kernel with a template size of 3*3 twice, and the number of channels is all 16, to obtain a feature map $\text{Map}_1$. Convolving an input of the current layer with a 5*5 convolution kernel to get the feature map $\text{Map}_2$, adding the $\text{Map}_1$ and the $\text{Map}_2$ to obtain a feature map $\text{Map}_3$, the above convolution step size is all 1, the number of channels of the output feature map is all 16, and finally stacking the $\text{Map}_1$, the $\text{Map}_2$ and the $\text{Map}_3$ to obtain the output of the current layer, where the number of channels is 48.

**Overall network structure.** The overall structure of gait age recognition is shown in Figure 5. The numbers in the figure denote the output channel size corresponding to the network layer. The whole system consists of five network parts. The first part is GLCL, which preliminarily extracts the features of global and local gait energy image and stacks them into 32-channel feature maps. The second, third, and fourth parts RE_a, RE_b, and RE_c are the residual expansion networks designed in this paper. RE_a is the residual expansion networks that cycles 3 times, and its output channel size is 32. RE_b is the residual expansion networks that cycles 5 times, and its output channel size is 64. RE_c is the residual expansion networks that cycles 5 times, and its output channel size is 128. The fifth part is mainly used for classification. For the output features of RE_c with 128 channels in total, five 1*1 convolution kernels are used to reduce the number of feature channels to 5. The main purpose is to keep the number of channels consistent with the number of classes, that is, the feature map of a channel represents a class, and the average pooling is adopted to replace the full connection to convert the feature map of five channels into a feature vector of length 5, and corresponding to five age groups. The average pooling can greatly reduce a large number of parameters and operations generated by using the full connection network, and finally the probability of converting the feature map into the corresponding class is converted by using Softmax. The maximum probability is selected as a prediction result of the network.
4. Experiments

4.1. Division of age groups
The dataset used in this paper is OULP-Age[18], which is the largest gait age dataset at present, including 63,846 pictures, and the size of each frame gait energy image is 128*88. The ages covered range from 2 to 90 years old. Figure 6 is a gait energy image for different age groups, from which it can be seen that there are certain differences in gait at different age groups. The distribution of the sample size in this dataset is extremely uneven, especially the sample size of the elderly is 2 orders of magnitude different from that of the young adults. This article divides age into 5 groups, which are 0-5, 6-10, 11-15, 16-60, and greater than 60. Among them, the largest number of samples are 16-60 years old. The overall network structure of deep learning used in the experiment is shown in Figure 5. P1-P4 are the maximum pooling layers, and stride is set to 2, which is used to compress the feature map, and P5 is the average pooling, which is used to replace the fully convolutional networks (FCN) layer and directly output a vector of the same length as the classification category for age group recognition. C1-C4 are all 1*1 convolutions, used to adjust the network channel size. The global local network structure is shown in Figure 3, and the residual expansion networks is shown in Figure 4.

Figure 6. Gait energy images of different age groups, (a) is the gait energy image of 0-6 years old, (b) is is the gait energy image of 6-10 years old, (c) is the gait energy image of 11-15 years old, (d) is the gait energy image of 16-60 years old, (e) is the gait energy image of over 60 years old.
4.2. Contrast test of weight restraint capability

According to the network structure shown in Figure 5, the FCN is trained using the training set until the network reaches the fitting state. The judgment basis of network fitting is when the training accuracy and the loss value tended to be stable. The gradient descent optimizer is selected as the optimization function for training the FCN. The loss function is the designed dynamic weight multi-classification loss function. A batch size is 256. When epoch is 16, the training is stopped to save the model. Automatic end-to-end identification can be achieved by feeding the test set into a saved model.

Figure 7 compares the loss values of the training process with the number of iterations under the three weight modes: fixed weight (G-W), dynamic weight (D-W) and no weight (N-W). The fixed weight value is obtained by calculating the proportion of each age group in the total sample set, which is [8.097, 1.323, 1.283, 0.312, 7.169]. The initial loss values are the same because the initialization parameters of the networks are consistent, but the loss value of the dynamic weight multi-classification cross-entropy loss function rapidly increases, and the loss value of the multi-classification cross-entropy loss function slowly decreases. This shows that the latter has already appeared a class in which neural networks tend to have multiple samples, under the influence of weights, the former enables neural networks to learn the features that represent classes with a small number of samples.

Table 1 shows the recognition rates of different age groups under the three patterns. The bold value represents the lowest recognition rate under each loss function. It is obvious that the proposed dynamic weight multi-classification cross-entropy loss function is superior to the fixed weight multi-classification cross-entropy loss function and the common multi-classification cross-entropy loss function. The overall recognition rate scores are 80.18%, 65.79%, and 76.72%. The unweighted loss function makes the FCN only learn the features of age groups with multiple samples, so that the features can not effectively represent a larger age range, while the dynamic weights make the features learned by the network better describe the differences among various age groups, thus effectively distinguishing different age groups.

Table 1

| Age Group | G-W | D-W | N-W |
|-----------|-----|-----|-----|
| 80-90     | 80% | 78% | 60% |
| 90-100    | 70% | 65% | 50% |
| 0-30      | 60% | 55% | 45% |

The mean absolute error (MAE) can reflect the prediction error, and the calculation formula is as follows:

\[ e_M = \frac{1}{N} \sum_{n=1}^{N} |T_n - P_n| \]  

Where \( N \) is the number of the entire test sample set, \( T \) is the true data label value, and \( P \) is the predicted value. Table 3 shows the MAE of dynamic weight, fixed weight and no weight, where the MAE of dynamic weight is the smallest, and the fixed weight is similar to no weight, which indicates that the prediction values of the FCN are more concentrated in the real categories under the constraint of dynamic weight.

Under the influence of the dynamic weights, the prediction results are gathered around the real age, and without the weights, the minority sample classes are obviously predicted as the majority sample class. The fixed weights have a certain constraint ability on the FCN, but are weaker than the dynamic weights. Since the proportion of the number of samples of each class in the training batch is correspondingly changed when the training batch is randomly divided, the dynamic weight can not only adapt to the change, but also realize greater constraint on the FCN according to the real-time prediction.

4.3. Feature descriptive power comparison test

In order to prove the effectiveness of the feature descriptive power extracted by the neural network under the dynamic weight constraint, it is compared with the HOG feature and the LBP feature. The latter two use KNN as the classifier. Table 2 shows the recognition accuracy rate (Acc rate) of each age group and the recognition accuracy rate of LBP feature and HOG feature, in the FCN under the dynamic weight constraint. The black bold is the lowest recognition rate. It can be seen that the depth features extracted by the FCN can better describe the differences between various age groups than the HOG feature and the LBP feature.
Figure 7. Iterative loss values of different weight patterns.

Table 1. Comparison of recognition rates of three weighted patterns.

| Age group | D-W | N-W | G-W |
|-----------|-----|-----|-----|
| 0-5       | 76.96 | 78.01 | 66.17 |
| 6-10      | 74.28 | 72.79 | 74.08 |
| 11-15     | 75.36 | 1.71 | 88.85 |
| 16-60     | 82.49 | 100.00 | 62.63 |
| >60       | **72.57** | **0.00** | **28.36** |
| Average   | 80.18 | 65.79 | 76.72 |

Figure 8. Comparison of the recognition accuracy rate of the three features.

Table 2. Comparison of the accuracy rate of the three features.

| Age group | D-W | HOG | LBP |
|-----------|-----|-----|-----|
| 0         | 76.96 | 35.10 | 21.35 |
| 1         | 74.28 | 64.58 | 29.89 |
| 2         | 75.36 | 25.42 | 4.82 |
| 3         | 82.49 | 94.61 | 94.90 |
| 4         | **72.57** | 6.53 | 1.68 |

Figure 7. Iterative loss values of different weight patterns.

Table 3. The mean absolute error of the three weights.

| Weight | D-W | N-W | G-W |
|--------|-----|-----|-----|
| MAE    | 0.20 | 0.36 | 0.35 |

Table 4. Average absolute error of three features.

| Feature | D-W | LBP | HOG |
|---------|-----|-----|-----|
| MAE     | 0.20 | 0.45 | 0.26 |

5. Conclusions
In this paper, inspired by the literature [17], we first decompose the human gait energy image into three parts: head-gait energy image, chest-gait energy image and leg-gait energy image, and design a global-local convolutional layer and combine the residual expansion module to fuse the global-local features of the gait energy image. Secondly, because the number of samples in each age group in the OULP-Age dataset is extremely uneven, a dynamic weight multi-classification cross-entropy loss function is designed. The experimental results show that the loss function can effectively solve the problem that the classification results are biased towards multiple samples. The comparison with the fixed weights
indicates that the dynamic weights can more effectively make the fully convolutional neural networks learn the more robust features. The comparison with HOG feature and LBP feature proves the descriptive power of the depth feature extracted by the fully convolutional networks under the dynamic weight constraint.

Acknowledgments
This work was supported in part by the Science and Technology Research Project of Chongqing Education Commission under Grant KJQN202005801 and KJQN201901123, in part by the Key Projects of Technological Innovation and Application Development of Chongqing Science and Technology Bureau under Grant No.cstc2019jscx-mbdxX0002.

References
[1] ANGULU R, TAPAMO J R, ADEWUMI A O. Age estimation via face images: a survey[J]. EURASIP Journal on Image and Video Processing, 2018, 2018(1): 42.
[2] Wang X, Liang L, Wang Z, Hu S. Research on age estimation technology of face images [J]. journal of image and graphics, 2012,17(06):603-618.
[3] Chen J, Li S, Gao Z, Li Z, Yang J. age and gender recognition based on improved CNN [J]. computer engineering and applications, 2018,54(16):135-139+175.
[4] MARTÍN-FELEZ R, XIANG T. Uncooperative gait recognition by learning to rank[J]. Pattern Recognition, 2014, 47(12): 3793-3806.
[5] BASHIR K, XIANG T, GONG S. Gait recognition without subject cooperation[J]. Pattern Recognition Letters, 2010, 31(13): 2052-2060.
[6] HAN J, BHANU B. Individual recognition using gait energy image[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2005, 28(2):316-322.
[7] TAKEMURA N, MAKIHARA Y, MURAMATSU D, et al. On input/output architectures for convolutional neural network-based cross-view gait recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017.
[8] GUI H, Liu J. Classification, diagnosis and treatment of gait disorders [J]. chinese journal of contemporary neurology and neurosurgery, 2011,11(01):18-21.
[9] MAKIHARA Y, OKUMURA M, IWAMA H, et al. Gait-based age estimation using a whole-generation gait database[C]//2011 International Joint Conference on Biometrics (IJCB). IEEE, 2011: 1-6.
[10] INCE O F, PARK J S, SONG J, et al. Child and adult classification using ratio of head and body heights in images[J]. International Journal of Computer and Communication Engineering, 2014, 3(2): 120.
[11] NABILA M, MOHAMMED A I, YOUSRA B J. Gait-based human age classification using a silhouette model[J]. IET Biometrics, 2017, 7(2): 116-124.
[12] Zhang C. Research on key technologies of gait recognition based on depth learning [D]. Beijing University of Posts and Telecommunications, 2016.
[13] He Y, Zhang J. Deep learning of gait recognition: a review [J]. Pattern Recognition and Artificial Intelligence, 2018,31(05):442-452.
[14] MARÍN-JIMÉNEZ M J, CASTRO F M, GUIL N, et al. Deep multi-task learning for gait-based biometrics[C]//2017 IEEE International Conference on Image Processing (ICIP). IEEE, 2017: 106-110.
[15] Yan J, Jiang Z, Liu L. Face gender recognition based on HOG- multiscale LBP feature [J]. Journal of Yangzhou University (Natural Science Edition), 2016,19(03):58-62.
[16] Cao L, Dikmen M, Fu Y, et al. Gender recognition from body[C]//Proceedings of the 16th ACM international conference on Multimedia. 2008: 725-728.
[17] Deng Y, Luo P, Loy C C, et al. Pedestrian attribute recognition at far distance[C]//Proceedings of the 22nd ACM international conference on Multimedia. 2014: 789-792.
[18] Liu P, Liu X, Yan J, et al. Localization Guided Learning for Pedestrian Attribute Recognition[C]//29th British Machine Vision Conference (BMVC 2018). BMVA Press, 2018.

[19] Zhao X, Sang L, Ding G, et al. Grouping Attribute Recognition for Pedestrian with Joint Recurrent Learning[C]//IJCAI. 2018: 3177-3183.

[20] M.D. ZEILER, RANZATOM, MONGA R, et al. On rectified linear units for speech processing[C]//Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013.

[21] A.L. MAAS, A.Y. HANNUN, A.Y. NG. Rectifier nonlinearities improve neural network acoustic models, in: Proceedings of the International Conference on Machine Learning (ICML), 30, 2013.

[22] ZHU H, ZHANG Y, LI G, et al. Ordinal Distribution Regression for Gait-based Age Estimation[J]. arXiv preprint arXiv:1905.11005, 2019.