Available online at www.sciencedirect.com

ScienceDirect

Journal of the Franklin Institute 000 (2017) 1–18

www.elsevier.com/locate/jfranklin

HOG-assisted deep feature learning for pedestrian gender recognition

Lei Cai[a], Jianqing Zhu[b], Huanqiang Zeng[a,*], Jing Chen[a], Canhui Cai[b], Kai-Kuang Ma[c]

[a] *School of Information Science and Engineering, Huaqiao University, Xiamen 361021, China*
[b] *School of Engineering, Huaqiao University, Quanzhou 362021, China*
[c] *School of Electrical and Electronic Engineering, Nanyang Technological University, 639798, Singapore*

Received 3 November 2016; received in revised form 6 July 2017; accepted 1 September 2017
Available online xxx

Abstract

Pedestrian gender recognition is a very challenging problem, since the viewpoint variations, illumination changes, occlusion, and poor quality are usually encountered in the pedestrian images. To address this problem, an effective *HOG-assisted deep feature learning* (HDFL) method is proposed in this paper. The key novelty lies in the design of HDFL network to effectively explore both deep-learned feature and weighted *histogram of oriented gradient* (HOG) feature for the pedestrian gender recognition. Specifically, the deep-learned and weighted HOG feature extraction branches are simultaneously performed on the input pedestrian image. A feature fusion process is subsequently conducted to obtain a more robust and discriminative feature, which is then fed to a softmax classifier for pedestrian gender recognition. Extensive experiments on multiple existing pedestrian image datasets have shown that the proposed HDFL method is able to effectively recognize the pedestrian gender, and consistently outperforms the state-of-the-art methods.

© 2017 The Franklin Institute. Published by Elsevier Ltd. All rights reserved.

[*] Corresponding author.
E-mail addresses: cailei_03@foxmail.com (L. Cai), jqzhu@hqu.edu.cn (J. Zhu), zeng0043@hqu.edu.cn (H. Zeng), jingzi@hqu.edu.cn (J. Chen), chcai@hqu.edu.cn (C. Cai), ekkma@ntu.edu.sg (K.-K. Ma).

http://dx.doi.org/10.1016/j.jfranklin.2017.09.003
0016-0032/© 2017 The Franklin Institute. Published by Elsevier Ltd. All rights reserved.

Please cite this article as: L. Cai et al., HOG-assisted deep feature learning for pedestrian gender recognition, Journal of the Franklin Institute (2017), http://dx.doi.org/10.1016/j.jfranklin.2017.09.003

1. Introduction

With the rapid development of internet, cloud computing, and multimedia technologies, digital video surveillance systems have been widely deployed in various areas, such as shopping mall, train station, airport, and so on. In these digital video surveillance systems, object identification plays a very important role for public safety [1]. Due to the extremely huge amount of visual data, various intelligent visual analytic tools have been developed as effective and essential solutions for identifying different attributes of objects in efficient and accurate manner, including visual object detection [2–6], face recognition [7], pedestrian re-identification [8,9], gender recognition [10], action recognition [11], race recognition [12], to name a few.

For a pedestrian, gender is one of the most important and useful attributes in many applications, e.g., human–computer interaction, video surveillance, health care, population statistics and multimedia retrieval system [13]. However, pedestrian gender recognition is a very challenge problem, which can be easily observed from some samples of pedestrian images selected from various datasets [14] as shown in Fig. 1. Firstly, in the practical scenarios, the appearances and postures of pedestrians are very diversified. Secondly, the pedestrian images captured by the surveillance cameras under different environments are often of poor quality, especially in the long distance condition. Specifically, the viewpoint variation, background clutter, illumination change and object occlusion are frequently encountered in the pedestrian images. Therefore, how to develop an effective pedestrian gender recognition method becomes a meaningful but difficult research topic in the computer vision field.

To address the problem of gender recognition, prior works generally follow the traditional image classification framework, which consists of two main stages: feature extraction and classification. The commonly used procedure is to design a highly representative and discriminative feature descriptor for gender in the first stage, followed by obtaining an accurate binary classifier that can well distinguish the difference between male and female in the second stage. Intuitively, a good feature representation should not only be robust to various noises but also be very discriminative. For that, many well-known hand-crafted features are developed based on the knowledge and expertise of the researchers and have been demonstrated their successes on the gender recognition based on the facial images, e.g., the *histogram of gradient* (HOG) [10], the *local binary pattern* (LBP) [15], iris code [16], etc.. However, they might not be effective for the gender recognition based on pedestrian images. This is because the pedestrian images are usually suffered from the view variations, occlusions, illumination changes, etc., and it is hard to obtain the reliable face information of the pedestrian in the real-world scenarios. Unlike the hand-crafted features, the *deep neural network* (DNN)-based methods can automatically learn effective representations of the input data so as to improve the performance of the classifier. Hence, instead of hand-crated features, the DNN, especially *convolutional neural network* (CNN), has achieved a great success in many computer vision tasks [17]. However, the DNN-based methods usually require the larger scale training dataset to learn an effective model.

Based on the above-mentioned analysis, the hand-crafted features would be more reliable to capture the local image characteristic while the deep-learned feature would be more adaptive to the input data. Intuitively, they could be considered as complementary parts to each other. In this paper, we propose an effective pedestrian gender recognition method by specially designing *HOG-assisted deep feature learning* (HDFL) network to effectively explore the advantages of both hand-crafted feature and deep-learned feature. To be more specific,

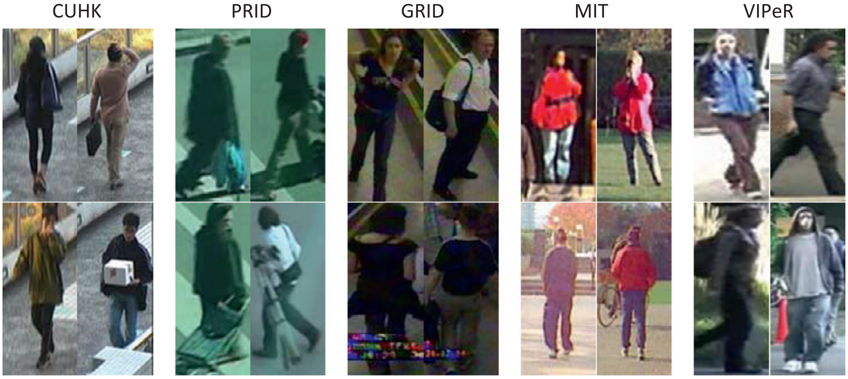

Fig. 1. Examples of pedestrian images selected from various pedestrian datasets.

the proposed HDFL approach simultaneously performs the deep-learned and weighted HOG feature extraction on the input pedestrian image in the first stage. The above-mentioned two features are then fused together as a more robust and discriminative feature for training a binary classifier for pedestrian gender recognition. Extensive experiments are carried out on multiple existing pedestrian datasets, CUHK, PRID, GRID, MIT, and VIPeR [14], showing that the proposed HDFL method outperforms the state-of-the-art pedestrian gender recognition methods.

The rest of this paper is organized as follows. Section 2 introduces the related gender recognition works. Section 3 describes the proposed pedestrian gender recognition method, HDFL, in detail. Section 4 presents the experimental results and discussions. Section 5 provides the conclusion.

2. Related work

In this section, we briefly review the existing gender recognition works, which can be roughly divided into two categorizes based on the types of input image: (1) face image-based gender recognition, and (2) pedestrian image-based gender recognition.

2.1. Face image-based gender recognition

Considering that facial feature is one of the most effective biometric characteristics for personal recognition, gender recognition based on face image is developed for those practical scenarios where are able to acquire clear face information. The existing methods extract different features to fully explore the face information for gender recognition from the holistic or local level, such as the geometric relationships between the facial landmarks [18], the raw pixel combination [19], pixel differences [20], Haar-like feature [21], LBP [15], and so on. Then, the extracted features are usually fed into a classifier for learning a discriminative model, for example, the AdaBoost and the *support vector machine* (SVM) classifiers have been widely employed [15,20,22].

In addition, some supervised learning methods, e.g., *Extreme learning machine* (ELM) [23,24] and *Deep Convolutional Neuron Network* (DCNN) [25,26], are also used for the face image-based gender recognition. For instance, Mahmood et al. [24] proposed a *fast adaptive shrinkage/thresholding algorithm ELM* (FASTA-ELM) for solving face image-based gender recognition problem. Levi and Hassncer [25] designed a DCNN to recognize the gender and age attributes based on the real-world face images. Mansanet et al. [26] presented a discriminative model, called *Local Deep Neural Network* (Local-DNN), to learn from the small overlapping regions in the visual field for gender recognition.

2.2. Pedestrian image-based gender recognition

In general, it is hard to obtain the clear face image in the practical sceneries. On the contrary, pedestrian images are easier to be captured at a distance without the pedestrian's cooperation. To this end, it could be more practical to develop the pedestrian image-based gender recognition. For that, some researchers have been devoted to infer the gender from the body structure of pedestrian. For example, gait feature, which describes the walking manner of a pedestrian, is extracted as an effective biometric characteristic for pedestrian gender recognition [27–30]. Yu et al. [27] used *gait energy image* (GEI) together with SVM. Lu and Tan [29] proposed a view-invariant gait-based gender recognition algorithm by using subspace learning. Hu et al. [28] presented a mixed conditional random field approach for gait-based gender recognition. Lu et al. [30] performed gender recognition based on pedestrian gait sequences with arbitrary walking directions. In addition, some traditional hand-crafted features, which are originally proposed for other object recognition problems, are applied on pedestrian gender recognition. Cao et al. [31] made the first attempt to employ the HOG feature and Adaboost classifier for exploiting silhouette information on gender recognition, in which each image was firstly divided into a collection of patches that model different parts of human body and further represented by HOG feature. Collins et al. [32] presented an improved HOG feature (namely, PixelHOG). The PixelHOG descriptor exploited the dense HOG features computed from an edge map and HSV color information based on the pixels' hue and saturation. Bourdev et al. [33] proposed a so-called *poselet* feature consisting of HOG, color histogram and skin, followed by training attribute classifiers using SVM to predict gender in unconstrained environments. Although the poselet feature is robust to pose variations and occlusion, it requires a lot of training data with detailed annotations of the human body. Guo et al. [34] employed *biologically-inspired feature* (BIF) derived from Gabor filters and a linear SVM classifier for gender classification.

In addition to the above-mentioned hand-crafted features, there are a few deep learning-based pedestrian gender recognition methods. Ng et al. [35] made the first attempt to train a CNN for gender recognition on the MIT dataset. Their CNN model involves two-stage convolution and subsampling layers and a fully-connected layer with 25 neuron units before prediction layer. Moreover, Ng et al. [36] further studied the image representation on different color spaces for training a CNN on gender recognition. Similar to the network structure in [35], Antipov et al. [37] trained a CNN model (denoted as Mini-CNN) and further fine-tuned a pre-trained CNN (called AlexNet) to effectively improve the pedestrian gender recognition rate.

3. HOG-assisted deep feature learning for pedestrian gender recognition

Fig. 2 shows the architecture of the proposed *HOG-assisted deep feature learning* (HDFL) Network for Pedestrian Gender Recognition. One can see that the proposed HDFL approach

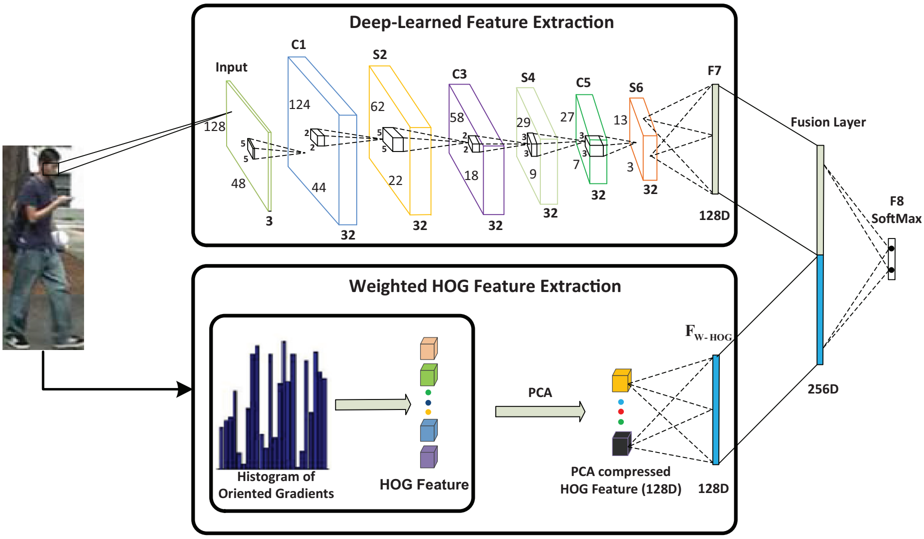
Deep-Learned Feature Extraction
Input
128
48
3
C1
124
44
32
S2
62
22
32
C3
58
18
32
S4
29
9
32
C5
27
7
32
S6
13
3
32
F7
128D
Fusion Layer
F8
SoftMax
256D
Weighted HOG Feature Extraction
Histogram of Oriented Gradients
HOG Feature
PCA
F_W-HOG
PCA compressed HOG Feature (128D)
128D

Fig. 2. Illustration of the framework of the Proposed HDFL.

starts with the deep-learned and weighted HOG feature extraction, followed by a feature fusion stage to obtain a more robust and discriminative feature. After that, a softmax classifier is learned on the fused feature for pedestrian gender recognition. The details of the proposed HDFL will be outlined in the following subsections.

3.1. Deep-learned feature extraction

There are many existing deeper and wider CNN models that have demonstrated their effectiveness on recognition problems, such as the VGGNet, GoogleNet, and so on. However, it is inappropriate to directly apply them on the pedestrian gender recognition task. This is because (1) they do not consider the special characteristic of gender attribute and pedestrian image, for example, pedestrian images are often captured under a long distance and of low resolutions; (2) they require a large scale training dataset to learn millions parameters of models, while the existing pedestrian image datasets are too small to train these deeper and wider modules. For that, a light convolutional neural network is adopted in the proposed HDFL to extract the deep-learned feature in this work. As shown in the upper part of Fig. 2, the deep-learned feature extraction branch consists of three convolution layers (i.e., C1, C3, C5), three subsampling layers (i.e., S2, S4, S6), and a fully connected layer with 128 neuron units (i.e., F7).

In the input layer, the input color images used in our HDFL approach are with the size of 48×128 and three channels (i.e., R, G, B). In the convolutional Cj (i.e., C1, C3, C5) layer, the feature maps are obtained by convolving a set of filters with output feature map of previous layer X_i, and then passing through *Batch Normalization* (BN) [38] and *rectified linear unit* (ReLU) activation function. First of all, the output feature map of convolution

Table 1
Algorithm steps of BN operation.

```
**Input:** Values of $y^k$ over a mini-batch: $\text{ß} = \{y_1^k, \dots, y_M^k\}$;
$y^k$ denotes the $k$–th element of feature map $Y_j$, and $M$ is the batch size.
**Output:** $v_i = BN_{\gamma,\beta}(y_i^k)$
$\mu_{\text{ß}} \leftarrow \frac{1}{M}\sum_{i=1}^{M} y_i^k$                     //mini-batch mean
$\sigma_{\text{ß}}^2 \leftarrow \frac{1}{M}\sum_{i=1}^{M}(y_i^k - \mu_{\text{ß}})^2$         // mini-batch variance
$\hat{y}_i \leftarrow \frac{y_i^k - \mu_{\text{ß}}}{\sqrt{\sigma_{\text{ß}}^2 + \epsilon}}$                 // normalization
$v_i \leftarrow \gamma\hat{y}_i + \beta \equiv BN_{\gamma,\beta}(y_i^k)$            //scale and shift
$\epsilon$ is a constant to ensure numerical stability.
$\gamma$ and $\beta$ are parameters needed to be learned.
```

operation Y_j in Cj layer can be formulated as

$$Y_j = \sum_{j \in N} W_{j,i} \otimes X_i + b_j \tag{1}$$

where $W_{j,\ i}$ means the weights of the filter with the size of $s_h \times s_w$, which connects the feature map of the previous layer X_i to the feature map Y_j, and b_j denotes the corresponding trainable bias, $\otimes$ denotes the convolution operation, N denotes the set of all or selected feature maps from previous layer. Assume that the size of an input feature map is $h \times w$, the convolutional layer with filters of size $s_h \times s_w$ and stride of d will produce a set of feature maps with a size of $(\frac{h-s_h}{d} + 1) \times (\frac{w-s_w}{d} + 1)$, disregarding the border effects. Then, the BN operation is applied on the output feature map of convolution operation Y_j, and the corresponding algorithm steps can be referred to Table 1. Finally, the output of BN, v_i, is further passed though the ReLU activation function f, which introduces non-linearities and is given as below:

$$f(v_i) = \max(0, v_i) \tag{2}$$

In the subsampling S_{j+1} (i.e., S2, S4, S6) layer, the feature maps are obtained by downsampling each feature map of the corresponding previous layer (i.e., C1, C3, C5). The downsampling is to exploit the maximum pooling operation. In other words, the largest value in a local region with a pre-set window size of previous layer is taken as the value of the feature map of the current layer. In the fully connected layer F7, each neuron unit fully connects to all the neuron units in the feature map of layer S6, in which the ReLU activation function is also employed and the *Local Response Normalization* (LRN) [39] is further used to perform normalization. Note that the LRN with across-channels region normalization is employed in this work, that is, in across-channels mode, the local regions extend across nearby channels. Specifically, let $a_{x,y}^i$ be the output value of a neuron unit i at position (x, y), the LRN output $b_{x,y}^i$ can be calculated as follows [39]:

$$b_{x,y}^i = \frac{a_{x,y}^i}{\left(k + \alpha \sum_{j=max(0,i-n/2)}^{min(N-1,i+n/2)} (a_{x,y}^j)^2\right)^{\beta}} \tag{3}$$

where the parameters, $k = 2$, $\alpha = 10^{-4}$, $\beta = 0.75$, $n = 5$, are set as suggested in [39], and $N = 128$ is equal to the number of neuron units in F7 layer of the proposed HDFL network.

In the proposed HDFL network, the number of filters in all convolutional layers is set to 32, since the size of pedestrian image dataset is small and less number of filters is beneficial

Please cite this article as: L. Cai et al., HOG-assisted deep feature learning for pedestrian gender recognition, Journal of the Franklin Institute (2017), http://dx.doi.org/10.1016/j.jfranklin.2017.09.003

Table 2
The parameter details of our specially designed HDFL network.

Layers	Types	Filter sizes	Filter number	Stride	Output sizes
C1	Input	–	–	–	3 × 128 × 48
	Convolution	5 × 5	32	1	32 × 124 × 44
	BN	–	–	–	32 × 124 × 44
	ReLU	–	–	–	32 × 124 × 44
S2	Maxpool	2 × 2	–	2	32 × 62 × 22
C3	Convolution	5 × 5	32	1	32 × 58 × 18
	BN	–	–	–	32 × 58 × 18
	ReLU	–	–	–	32 × 58 × 18
S4	Maxpool	2 × 2	–	2	32 × 29 × 9
C5	Convolution	3 × 3	32	1	32 × 27 × 7
	BN	–	–	–	32 × 27 × 7
	ReLU	–	–	–	32 × 27 × 7
S6	Maxpool	3 × 3	–	2	32 × 13 × 3
F7	FC	1 × 1	128	–	128 × 1 × 1
	ReLU	–	–	–	128 × 1 × 1
	LRN	–	–	–	128 × 1 × 1
F_{W--HOG}	FC	1 × 1	128	–	128 × 1 × 1
	BN	–	–	–	128 × 1 × 1
	ReLU	–	–	–	128 × 1 × 1
	LRN	–	–	–	128 × 1 × 1
Fusion	Concat	–	–	–	256 × 1 × 1
F8	FC	1 × 1	2	–	2 × 1 × 1

to avoid over-fitting problem. In addition, the tiny-sized filters, i.e., 5 × 5, 5 × 5, and 3 × 3, are individually applied in C1, C3, and C5 layer for saving the filter parameters. This is because the filters with smaller sizes are more efficient to extract the deeper image details and thus more appropriate for pedestrian modelling, as the pedestrian usually has rich texture information. The stride of all the convolutional layers is set to 1 for retaining the image details as much as possible. For S2, S4, S6 layers, the 2 × 2, 2 × 2, 3 × 3 max pooling operation is used, respectively, and the stride is all set to 2. More parameter configuration of the deep-learned feature extraction branch in our designed HDFL network can be referred to Table 2.

3.2. Weighted HOG feature extraction

Many hand-crafted features are commonly used for various computer vision tasks [15,21,31,40], which could be considered as a complementary part to the deep-learned feature. Based on this intuition, the *histogram of oriented gradient* (HOG) [40] feature is employed to assist the deep feature learning for recognizing the pedestrian gender, since the HOG feature can effectively describe the local contour of input image and is robust to the illumination change, orientation variation, etc.. Specifically, as shown in the lower part of Fig. 2, a weighted HOG feature extraction branch is developed in the proposed HDFL network, which can be described as below:

Please cite this article as: L. Cai et al., HOG-assisted deep feature learning for pedestrian gender recognition, Journal of the Franklin Institute (2017), http://dx.doi.org/10.1016/j.jfranklin.2017.09.003

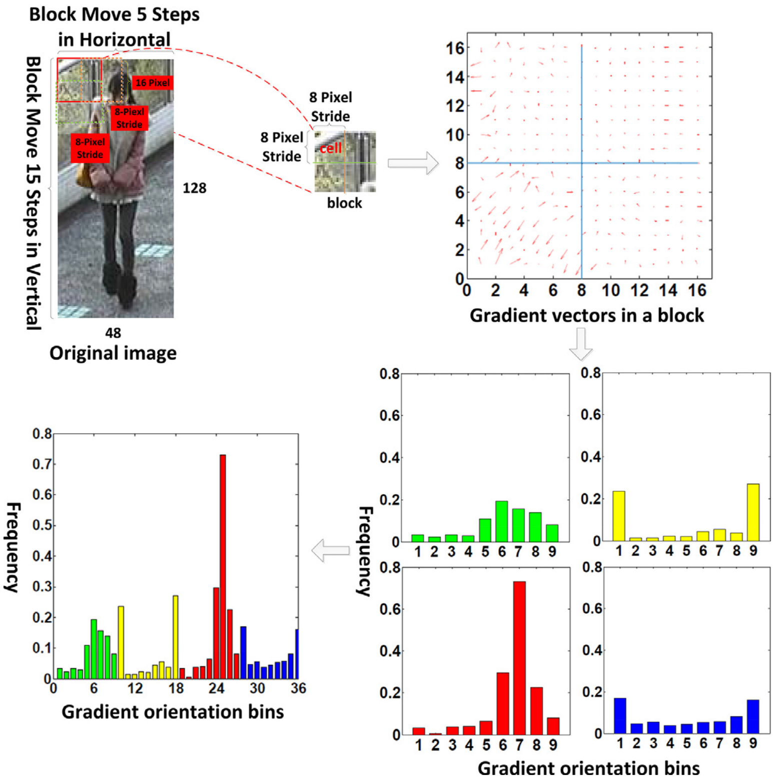

Fig. 3. The computational process of HOG feature.

(1) Gradient magnitude and orientation computation: for the input pedestrian image, the gradient of each pixel can be computed in the first stage:

$$G_h(x, y) = I(x+1, y) - I(x-1, y)$$
$$G_v(x, y) = I(x, y+1) - I(x, y-1) \tag{4}$$

where $I(x, y)$ denotes the pixel value at location (x, y) in an input pedestrian image, $G_h(x, y)$ and $G_v(x, y)$ mean the horizontal and vertical gradient, respectively. Then, the gradient magnitude $G(x, y)$ and gradient orientation $\alpha(x, y)$ can be calculated as below:

$$G(x, y) = \sqrt{G_h(x, y)^2 + G_v(x, y)^2}$$
$$\alpha(x, y) = \arctan \frac{G_v(x, y)}{G_h(x, y)} \tag{5}$$

(2) HOG feature vector generation: as shown in Fig. 3, the input image (i.e., 128 × 48) is firstly divided into 15 × 5 blocks with the size of 16 × 16 using the overlapping strategy via a 8-pixel stride, since the overlapping strategy between neighboring blocks can enhance the local correlation. Each block will be further divided into 4 cells with the size of 8 × 8. For each cell, nine Bins splitted over $0 - \pi$ are used

Please cite this article as: L. Cai et al., HOG-assisted deep feature learning for pedestrian gender recognition, Journal of the Franklin Institute (2017), http://dx.doi.org/10.1016/j.jfranklin.2017.09.003

to accumulate the gradient magnitude on the corresponding gradient direction. Specifically, let $\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_9$ individually present $0, \pi/9, 2\pi/9, \ldots, \pi$, For each pixel (x, y) in the cell, if $\alpha_i \leq \alpha(x, y) < \alpha_{i+1}$ $(i \in 0, 1, 2, \ldots, 8)$, then Bin B_i is chosen and $B_i = B_i + G(x, y)$. To have a better robustness to illumination change and noise, a L2-norm normalization step is performed on the obtained Bin B_i as follows:

$$B_i' = \frac{B_i}{\sqrt{\|B_i\|_2^2 + \phi}} \tag{6}$$

where ϕ is a small positive value to ensure the numerical stability. Hence, it will produce a 9-dimensional gradient orientation histogram vector B_i' for each cell, and consequently a $4 \times 9 = 36$ dimensional HOG feature vector for each block, as illustrated in Fig. 3.

(3) *Weighted HOG* (W-HOG) feature via fully connection: to adopt the HOG feature in our specially designed HDFL network, a fully connected layer $\mathrm{F}_{W\text{-}HOG}$ with 128 neuron units is employed to fully connect the HOG feature. However, the obtained HOG feature vector by step (2) is with the dimension $36 \times 5 \times 15 = 2700$ for an input pedestrian image, which would lead to a lot of parameters. To reduce the number of parameters in the fully-connected layer $\mathrm{F}_{W\text{-}HOG}$, the traditional *Principal Component Analysis* (PCA) [41] is performed to reduce the HOG feature dimension from 2700D to 128D. After that, the PCA compressed HOG feature is fully connected to the layer $\mathrm{F}_{W\text{-}HOG}$, in which the BN, ReLU activation function, and the LRN are used. More parameter configuration of the weighted HOG feature extraction branch in our designed HDFL network is listed in Table 2.

3.3. Deep-learned and weighted HOG feature fusion

Considering that the weighted HOG feature could effectively assist the deep-learned feature, the proposed HDFL network exploited a Feature Fusion Layer to combine these two kinds of features to produce a more robust and discriminative feature (i.e., *HOG-assisted deep-learned feature*, HDF) for pedestrian gender recognition, as illustrated in Fig. 2. Consequently, the HDF feature with 256 dimension, $\mathbf{a}_{HDF}$, can be presented as

$$\mathbf{a}_{HDF} = [\mathbf{a}_{DF}, \mathbf{a}_{W\text{-}HOG}] = [a_{0,DF}, \ldots, a_{127,DF}, a_{128,W-HOG}, \ldots, a_{255,W\text{-}HOG}] \tag{7}$$

where the $\mathbf{a}_{DF}$ and $\mathbf{a}_{W\text{-}HOG}$ denote the deep-learned feature vector and weighted HOG feature vector, respectively.

Through such feature fusion processes, the proposed HDFL method could learn a model to effectively explore the merits of both deep-learned and weighted HOG features for pedestrian gender recognition. This can be observed from both forward and back propagation processes, as follows. In the forward propagation process, the prediction score in the layer F8, $s(\mathbf{a}_{HDF})$, can be computed according to the forward propagation algorithm:

$$s(\mathbf{a}_{HDF}) = W^T \mathbf{a}_{HDF} + b \tag{8}$$

where W and b denote the weights and bias term, respectively. They are used to project the HDF feature $\mathbf{a}_{HDF}$ into the prediction score. The obtained prediction score $s(\mathbf{a}_{HDF})$ is then fed to a 2-way softmax for calculating the loss. The softmax loss can be minimized in terms of cross-entropy loss:

$$L = -\sum_{j=1}^{2} p_j \log p_j. \tag{9}$$

where p means the output of likelihood of the softmax, and can be formulated as

$$p_i(s) = \frac{exp(s_i)}{\sum_{j=1}^{2} exp(s_j)} \tag{10}$$

From the above equations, it can be clearly observed that the softmax loss L will be influenced by both deep-learned feature vector $\mathbf{a}_{DF}$ and weighted HOG feature vector $\mathbf{a}_{W\text{-}HOG}$ in the forward propagation process.

Besides, in the back propagation process, our proposed HDFL method aims to minimize the objective function J [42], which is solved by using *stochastic gradient decent* (SGD). Let W_{ij}^m be the weight connecting the j–th unit in $(m-1)$–th layer and the i–th unit in m–th layer, and $z_i^m = \sum_i W_{ij}^m a_j^{m-1}$, where $a_j^{m-1} = f(z_j^{m-1})$ and f denotes the activation function. Taking the layer F7 as an example, we obtain

$$\frac{\partial J}{\partial W_{ij}^7} = x_{j,DF}^6 \delta_i^7 \tag{11}$$

where $x_{j,DF}^6$ is j–th element of output feature map of layer S6, and

$$\delta_i^7 = \left(\sum_k W_{ki}^8 \delta_k^8\right) f'(z_{i,DF}^7) \tag{12}$$

$$\delta_k^8 = \frac{\partial J}{\partial z_k^8} \tag{13}$$

It should be pointed out that $z_k^8 = \sum_k W_{ki}^8 a_{i,HDF}$, where $a_{i,\ HDF}$ is the ith element of the HDF feature vector $\mathbf{a}_{HDF}$ as shown in (7). Therefore, it can be seen that through $a_{i,HDF} \rightarrow z_k^8 \rightarrow \delta_k^8 \rightarrow \delta_i^7 \rightarrow \frac{\partial J}{\partial W_{ij}^7}$, the $\frac{\partial J}{\partial W_{ij}^7}$ will be affected by both deep-learned feature vector $\mathbf{a}_{DF}$ and weighted HOG feature vector $\mathbf{a}_{W\text{-}HOG}$ in the back propagation process.

3.4. Implementation details

Some implementation details of the proposed HDFL method can be described as follows. Firstly, the proposed HFDL model is trained by using *stochastic gradient decent* (SGD) [43], where the initial learning rate is set as $l = 0.01$ and decreased by $l \times 0.1$ after every 1500 iterations. All images used in our experiments are resized to 48 × 128 and subtracted their corresponding mean values. In order to avoid over-fitting problem, the horizontal mirrored copies of the training images are used to augment the training data, and the overall training data are randomly shuffled. During the training phase, the weights are initialized from a normal distribution $N(0, 0.01)$ and the biases are initialized as 0. In each iteration, a batch size of 128 samples is fed to the HDFL network for training, and the weights are updated by using back propagation. The details of the parameter configuration of the proposed HDFL network are shown in Table 2.

Table 3
Pedestrian datasets.

Datasets	Image numbers	Resolution	Environment
CUHK	4563	80 × 160	Outdoor (camera in high angle)
PRID	1134	64 × 128	Outdoor (most profile view)
GRID	500 + 777 (background)	From 29 × 67 to 169 × 365	Underground station (8 disjoint camera views)
MIT	888	64 × 128	Frontal (420) and rear (468) views
VIPeR	1264	48 × 128	Outdoor

Table 4
Training and testing images from each dataset.

Datasets	Training size (♂ + ♀)	Testing size (♂ + ♀)
CUHK	3844 = (2715+1129)	379 = (190+189)
PRID	947 = (458+489)	101 = (50+51)
GRID	928 = (531+397)	100 = (50+50)
MIT	788 = (538+250)	84 = (42+42)
VIPeR	1113 = (546+567)	120 = (60+60)
Total	7620	784

4. Experimental results and discussions

4.1. Datasets and evaluation protocol

In this section, the performances resulted from the proposed HDFL method and other state-of-the-art methods are compared based on multiple widely-used and challenging datasets [14], including CUHK, PRID, GRID, MIT, and VIPeR. These datasets as shown in Table 3 contain different kinds of pedestrian images, largely varying in appearances of the pedestrian, resolutions, environments (indoors or outdoors), and camera viewpoints (profile, frontal, rear, and high angle, etc.). By following the same practice in [37], we also filter out some images that consist of the same pedestrian or unidentified target or are with very low resolution. The corresponding numbers of training and testing images from each dataset are shown in Table 4. Our experiments randomly divide 8404 images from all the datasets into 2 parts: 7620 images for training and 784 images for testing, and repeats the random trials 10 times to obtain the average *Mean Average Precision* (MAP) and *Area Under ROC Curve* (AUC) [44] as the final results.

4.2. Performance comparison

To demonstrate the superiority, the proposed HDFL method is compared with some existing deeper networks and state-of-the-art pedestrian gender recognition methods, including Mini-CNN [37], AlexNet-CNN [37], VGGNet16 [45], GoogleNet [46], ResNet50 [47]. To further analyze how much of the contributions coming from the deep-learned feature and weighted HOG feature respectively, the performances resulted from *deep feature learning* (DFL) solely and *Weighted HOG* (W-HOG) solely are also investigated. Note that the DFL network and weighted HOG are exactly the deep-learned feature extraction branch and weighted HOG feature extraction branch in Fig. 2, respectively, except that 256 neuron units are employed

Table 5
Performance comparison.

Methods	MAP	AUC
Mini-CNN [37]	0.80	0.88
AlexNet-CNN [37]	0.85	0.91
VGGNet16 [45]	0.87	0.89
GoogleNet [46]	0.90	0.91
ResNet50 [47]	0.89	0.90
Proposed W-HOG	0.85	0.86
Proposed DFL	0.92	0.93
Proposed HDFL	**0.94**	**0.95**

Table 6
Comparison between the DF and HDF in terms of distance of centroid.

Feature	DF	HDFL
D_C	9.2	10.03

in layers F7 and $F_{W\text{-}HOG}$ for having a fair comparison to the proposed HDFL that has 256 neuron units in the feature fusion layer.

Table 5 shows the performances of these methods on the datasets documented in Table 3. One can clearly see that the proposed HDFL method is able to achieve the highest MAP and AUC, and consistently outperforms the state-of-the-art methods under comparison. In addition, it can be further observed that the DFL and W-HOG achieve relatively good performances, and the proposed HDFL that jointly explores the deep-learned and weighted HOG features performs the best. This investigation suggests that the deep-learned feature and weighted HOG feature effectively convey different characteristics of pedestrian gender, and they play a complementary role to jointly capture the gender properties well for delivering accurate pedestrian gender recognition.

4.3. Analysis of the proposed HDFL method

4.3.1. Assistance behavior of weighted HOG feature to deep-learned feature

To analyze the assistance behavior of weighted HOG feature to deep-learned feature, the extracted *deep-learned feature* (DF) and the *HOG-assisted deep-learned feature* (HDF) are visualized in Fig. 4. Moreover, the well-known squared Euclidean distance of centroid D_C [48] is exploited to quantitatively measure the discriminative ability of the DF and HDF, which can be defined as follows:

$$D_C = \|\overline{X}_F - \overline{X}_M\|^2 \tag{14}$$

where $\overline{X}_F$ and $\overline{X}_M$ denote the mean feature vectors of the female and male classes, respectively. Note that the z-score normalization is performed on the DF and HDF to make the extracted DF and HDF at the same scale, and the mean feature vectors ($\overline{X_F}$ and $\overline{X_M}$) are calculated based on the normalized DF and HDF. Note that a larger value of D_C indicates a larger inter-class distance. The comparison between the DF and HDF in terms of distance of centroid is shown in Table 6.

One can see from Fig. 4 that the intersection area of the scatter plot of the HDF is smaller than that of the DF. In other words, the HDF has a better separability than the DF. Moreover,

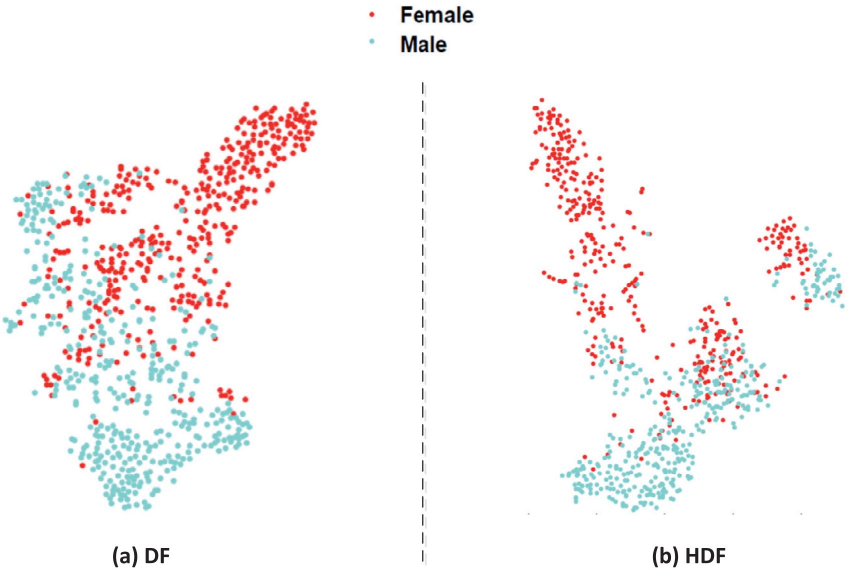

Fig. 4. Visualization of *deep-learned feature* (DF) and *HOG-assisted deep-learned feature* (HDF).

Table 7
Performance of the proposed HDFL without and with LRN.

Methods	MAP	AUC
HDFL without LRN	0.93	0.94
HDFL with LRN	0.94	0.95

it can be further observed from Table 6 that the HDF yields a larger distance of centroid (i.e., larger inter-class distance), compared with that of the DF. All these clearly indicate that the HDF is more discriminative than the DF. And the proposed HDFL is thus able to achieve higher recognition rates than the proposed DFL as documented in Table 5. This is due to the effective assistance behavior of weighted HOG feature to DF.

4.3.2. Effectiveness of local response normalization (LRN)

To demonstrate the effectiveness of LRN, experiments have been conducted to compare the performance of the proposed HDFL without and with LRN. The corresponding results are documented in Table 7. One can see from Table 7 that LRN improves the performance by 0.01 MAP and 0.01 AUC, and is useful for the proposed HDFL. To better explain the underlying cause, Fig. 5 provides an example of HDF feature vector resulted from the proposed HDFL without and with LRN. It can be observed that the LRN can normalize the deep-learned feature and weighted HOG feature into the similar scale. This would be very beneficial to feature fusion, in which the weighted HOG feature could be more effective to assistant the deep-learned feature for producing a more discriminative HDF.

4.3.3. Misclassified samples analysis

In the pedestrian gender recognition problem, misclassification (i.e., the pedestrian gender is male and wrongly classified as female, and vice versa) is unavoidable. To conduct a

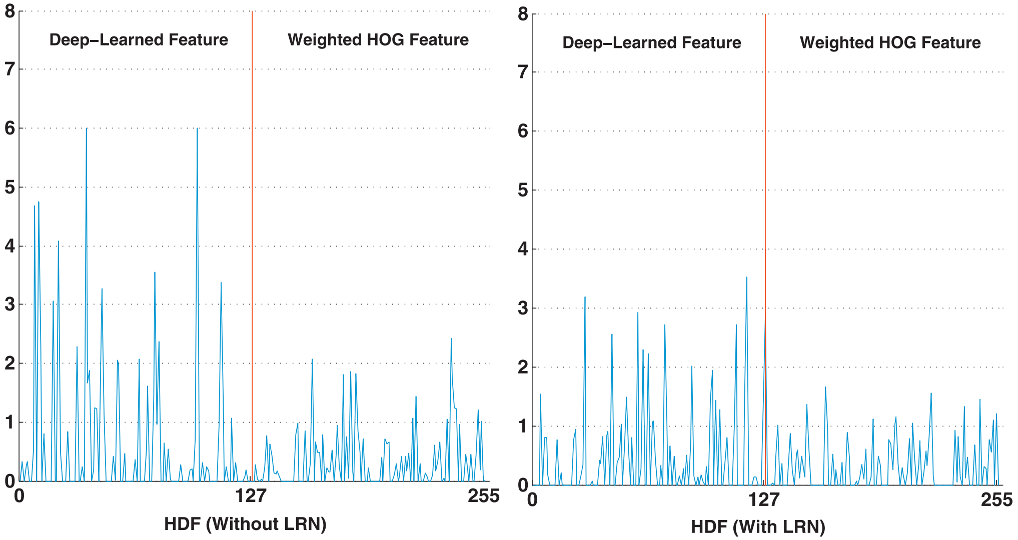

Fig. 5. HDF resulted from proposed HDFL without and with LRN.

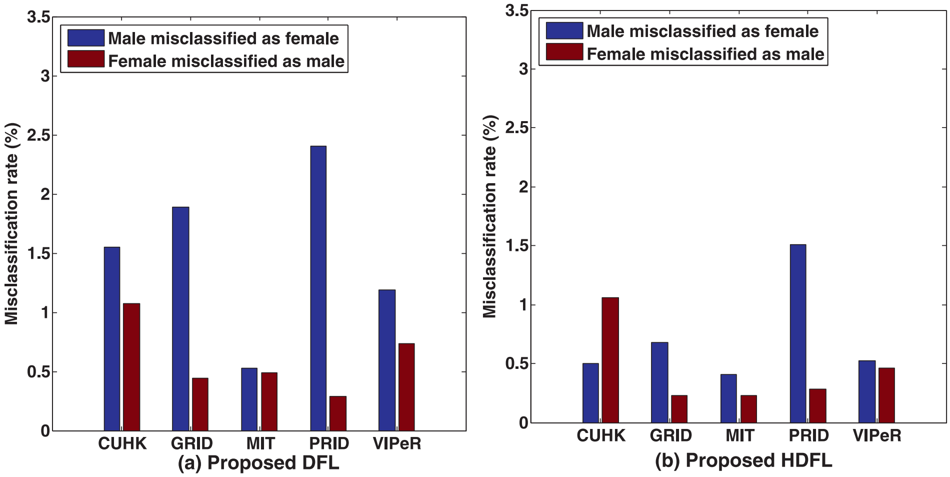

Fig. 6. Misclassification rate on each dataset.

comprehensively misclassified samples analysis, experiments are performed by using the proposed DFL and HDFL on all the images in the datasets. Fig. 6 shows the misclassification rate in various datasets. For each dataset, the misclassification rate is computed as the ratio between the number of the misclassified images and the total images. It can be found that the proposed HDFL significantly reduces the misclassification rate, compared with the proposed DFL. This further demonstrates the observation that the HDF is more distinctive than DF due to the assistance behavior of weighted HOG feature.

Moreover, Fig. 7(a) shows some examples of pedestrian images that are misclassified by the proposed DFL while correctly classified by the proposed HDFL, and Fig. 7(b) provides

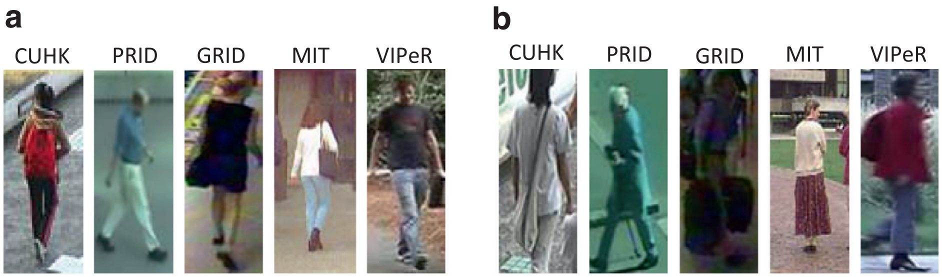

Fig. 7. Examples of misclassified pedestrian images in various datasets: (a) misclassification only by the proposed DFL; (b) misclassification by both proposed DFL and HDFL.

Table 8
Testing images for cross-dataset evaluation.

Datasets	Testing size (♂ + ♀)
3DPeS	100 = (50+50)
CAVIAP	68 = (34+34)
i-LIDS	100 = (50+50)
SARC3D	40 = (20+20)
TownCentre	42 = (21+21)
Total	350

some examples of pedestrian images misclassified by both proposed DFL and HDFL methods. It can be seen that the pedestrian images in Fig. 7(b) are more challenging than Fig. 7(a) overall. For example, in the 1st image of Fig. 7(b), the male with long hair is easier to be misclassified, since the long hair is one of the major features of female. Besides, although the 2nd and 5th images in Fig. 7(b) are female, they are hard to be distinguished, even by a human observer. In addition, the incomplete silhouette (e.g., the 1st image in Fig. 7(b)), occlusion, distortion (e.g., the 3rd image in Fig. 7(b)), and unusual pose (e.g., the 4th image in Fig. 7(b)), etc., are very difficult for pedestrian gender recognition problem, leading to misclassification. Future work will include devising a better model to improve the pedestrian recognition rate by further collecting more training images and considering more discriminative hand-crafted feature based on the proposed HDFL network.

4.4. Cross-dataset evaluation

To demonstrate the generalization ability, the proposed HDFL method, some existing deeper networks and state-of-the-art pedestrian gender recognition methods, including Mini-CNN [37], AlexNet-CNN [37], VGGNet16 [45], GoogleNet [46], ResNet50 [47], are tested on multiple completely unseen datasets [14], i.e., 3DPeS, CAVIAR, i-LIDS, SARC3D and TownCentre. In this cross-dataset evaluation, the numbers of testing pedestrian images randomly selected from each unseen dataset are shown in Table 8. This experiment also repeats the random trials 10 times to obtain the average MAP and AUC as the final results. The corresponding results are shown in Table 9.

It is interesting to see that the performances of all the methods are decreased, compared with that in Table 5. This is because the testing images are completely unseen for these methods in this cross-dataset evaluation. It can be further observed that the proposed HDFL

Please cite this article as: L. Cai et al., HOG-assisted deep feature learning for pedestrian gender recognition, Journal of the Franklin Institute (2017), http://dx.doi.org/10.1016/j.jfranklin.2017.09.003

Table 9
Performance comparison in cross-dataset evaluation.

Methods	MAP	AUC
Mini-CNN [37]	0.75	0.80
AlexNet-CNN [37]	0.79	0.85
VGGNet16 [45]	0.83	0.84
GoogleNet [46]	0.83	0.83
ResNet50 [47]	0.85	0.86
Proposed W-HOG	0.79	0.79
Proposed DFL	0.87	0.89
Proposed HDFL	**0.89**	**0.91**

also achieves the highest MAP and AUC, and is consistently superior to the state-of-the-art methods under comparison. This cross-dataset evaluation reveals that the proposed HDFL method has good generalization ability.

5. Conclusion

In this paper, a novel *HOG-assisted deep feature learning* (HDFL) method is proposed for pedestrian gender recognition. The superior performance is achieved by designing a special HDFL network to fully explore both deep-learned feature and weighted HOG feature. By designing a special HDFL network, the deep-learned and weighted HOG features are simultaneously extracted for the input pedestrian image and then fused together to obtain a more discriminative feature, which is then fed into the softmax classifier for pedestrian gender recognition. Extensive experiments on multiple challenging datasets show the proposed HDFL method can effectively recognize the pedestrian gender, yield good generalization ability, and consistently outperform the state-of-the-art methods.

Acknowledgments

This work was supported by the National Natural Science Foundation of China [grant numbers 61401167, 61372107, and 61602191]; the Natural Science Foundation of Fujian Province [grant numbers 2016J01308 and 2017J05103]; the Promotion Program for Young and Middle-aged Teacher in Science and Technology Research of Huaqiao University [grant numbers ZQN-YX403 and ZQN-PY418]; the Opening Project of State Key Laboratory of Digital Publishing Technology [grant number FZDP2015-B-001]; the High-Level Talent Project Foundation of Huaqiao University [grant numbers 14BS201, 14BS204 and 16BS108]; and the Graduate Student Scientific Research Innovation Project Foundation of Huaqiao University.

References

[1] A. Filonenko, K.H. Jo, Unattended object identification for intelligent surveillance systems using sequence of dual background difference, IEEE Trans. Indust. Inf. 12(6) (2016) 2247–2255.
[2] P. Felzenszwalb, R. Girshick, D. McAllester, Cascade object detection with deformable part models, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2010, pp. 2241–2248.
[3] R. Girshick, F. Iandola, T. Darrell, Deformable part models are convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 437–446.
[4] P. Felzenszwalb, R. Girshick, D. McAllester, Visual object detection with deformable part models, Commun. ACM 56(9) (2013) 97–105.

[5] R. Girshick, Fast R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
[6] J. Cao, W. Wang, J. Wang, R. Wang, Excavation equipment recognition based on novel acoustic statistical features, IEEE Trans. Cybern. (2016), doi:10.1109/TCYB.2016.2609999.
[7] Q. Feng, C. Yuan, J.S. Pan, J.F. Yang, Y.T. Chou, Y. Zhou, W. Li, Superimposed sparse parameter classifiers for face recognition, IEEE Trans. Cybern. 47(2) (2017) 378–390.
[8] T. Xiao, H. Li, W. Ouyang, Learning deep feature representations with domain guided dropout for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1249–1258.
[9] S. Wu, Y.C. Chen, X. Li, A.C. Wu, J.J. You, W.S. Zheng, An enhanced deep feature representation for person re-identification, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2016, pp. 1–8.
[10] J.E. Tapia, C.A. Perez, Gender classification based on fusion of different spatial scale features selected by mutual information from histogram of lbp, intensity, and shape, IEEE Trans. Inf. Forensics Secur. 8(3) (2013) 488–499.
[11] Y. Kong, Z. Ding, J. Li, Y. Fu, Deeply learned view-invariant features for cross-view action recognition, IEEE Trans. Image Process. 26(6) (2017) 2247–2255.
[12] S. Fu, H. He, Z.G. Hou, Learning race from face: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 36(12) (2013) 2483–2509.
[13] C.B. Ng, Y.H. Tay, B.M. Goi, Vision-based human gender recognition: a survey, arXiv preprint, 2013, arXiv:1204.1611.
[14] Y. Deng, P. Luo, C.C. Loy, X. Tang, Pedestrian attribute recognition at far distance, in: Proceedings of the Twenty-first ACM International conference on Multimedia, 2014, pp. 789–792.
[15] C.F. Shan, Learning local binary patterns for gender classification on real-world face images, Pattern Recogn. Lett. 33 (2012) 431–437.
[16] J.E. Tapia, C.A. Perez, K.W. Bowyer, Gender classification from the same iris code used for recognition, IEEE Trans. Inf. Forensics Secur. 11(8) (2016) 1760–1770.
[17] J. Lemley, S. Bazrafkan, P. Corcoran, Deep learning for consumer devices and services: pushing the limits for machine learning, artificial intelligence, and computer vision, IEEE Consum. Electron. Mag. 6(2) (2017) 48–56.
[18] C. BenAbdelkader, P. Griffin, A local region-based approach to gender classification from face images, in: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, p. 52.
[19] B. Moghaddam, M.H. Yang, Learning gender with support faces, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002) 707–711.
[20] S. Baluja, H.A. Rowley, Boosting sex identification performance, Int. J. Comput. Vis. 71 (2007) 111–119.
[21] G. Shakhnarovich, P.A. Viola, B. Moghaddam, A unified learning framework for real time face detection and classification view document, in: Proceedings of Fifth IEEE International Conference on Automatic Face and Gesture Recognition, 2002, pp. 14–21.
[22] E. Eidinger, R. Enbar, T. Hassner, Age and gender estimation of unfiltered faces, IEEE Trans. Inf. Forensics Secur. 9(12) (2014) 2170–2179.
[23] J. Cao, K. Zhang, M. Luo, C. Yin, X. Lai, Extreme learning machine and adaptive sparse representation for image classification, Neural Netw. 81 (2016) 91–102.
[24] S.F. Mahmood, M.H. Marhaban, F.Z. Rokhani, Fasta-elm: a fast adaptive shrinkage/thresholding algorithm for extreme learning machine and its application to gender recognition, Neurocomputing 219(5) (2017) 312–322.
[25] G. Levi, T. Hassncer, Age and gender classification using convolutional neural networks, in: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015, pp. 34–42.
[26] J. Mansanet, A. Albiol, R. Paredes, Local deep neural networks for gender recognition, Pattern Recogn. Lett. 70(15) (2016) 80–86.
[27] S. Yu, T. Tan, K. Huang, K. Jia, X. Wu, A study on gait-based gender classification, IEEE Trans. Image Process. 18(8) (2009) 1905–1910.
[28] M. Hu, Y. Wang, Z. Zhang, D. Zhang, Gait-based gender classification using mixed conditional random field, IEEE Trans. Syst. Man Cybern. Part B Cybern 41(5) (2011) 1429–1439.
[29] J. Lu, Y.P. Tan, Uncorrelated discriminant simplex analysis for view-invariant gait signal computing, Pattern Recogn. Lett. 31(5) (2010) 382–393.
[30] J. Lu, G. Wang, P. Moulin, Human identity and gender recognition from gait sequences with arbitrary walking directions, IEEE Trans. Inf. Forensics Secur. 9(1) (2014) 51–61.

[31] L. Cao, M. Dikmen, Y. Fu, T.S. Huang, Gender recognition from body, in: Proceedings of the Sixteeth ACM International Conference on Multimedia, 2008, pp. 725–728.

[32] M. Collins, J. Zhang, P. Miller, H. Wang, Full body image feature representations for gender profiling, in: Proceedings of the Twefth IEEE International Conference on Computer Vision Workshops, 2009, pp. 1235–1242.

[33] L. Bourdev, S. Malik, J. Malik, Describing people: a poselet-based approach to attribute classification, in: Proceedings of the 2011 IEEE International Conference on Computer Vision, 2011, pp. 1543–1550.

[34] G. Guo, G. Mu, Y. Fu, Gender from body: a biologically-inspired approach with manifold learning, in: Proceedings of the Asian Conference on Computer Vision, 2009, pp. 236–245.

[35] C.B. Ng, Y.H. Tay, B.M. Goi, A convolutional neural network for pedestrian gender recognition, in: Proceedings of the International Symposium on Neural Networks, 2013, pp. 558–564.

[36] C.B. Ng, Y.H. Tay, B.M. Goi, Comparing image representations for training a convolutional neural network to classify gender, in: Proceedings of the First International Conference on Artificial Intelligence, Modelling, 2013, pp. 29–33.

[37] G. Antipov, S.A. Berrani, N. Ruchaud, J.L. Dugelay, Learned vs. hand-crafted features for pedestrian gender recognition, in: Proceedings of the Twenty-thirth ACM International conference on Multimedia, 2015, pp. 1263–1266.

[38] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: Proceeding of the International Conference on Machine Learning, 2015, pp. 448–456.

[39] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[40] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, pp. 886–893.

[41] H. Abdi, L.J. Williams, Principal component analysis, Wiley Interdiscipl. Rev. Comput. Stat. 2(4) (2010) 433–459.

[42] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.

[43] L. Bottou, Stochastic gradient descent tricks, Neural Netw. Tricks Trade (2012) 421–436.

[44] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143(1) (1982) 29–36.

[45] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint, 2014, arXiv: 1409.1556.

[46] C. Szegedy, W. Liu, Y. Jia, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[47] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[48] J. Wu, H. Xiong, C. Liu, J. Chen, A generalization of distance functions for fuzzy c-means clustering with centroids of arithmetic means, IEEE Trans. Fuzzy Syst. 20(3) (2012) 557–571.

Please cite this article as: L. Cai et al., HOG-assisted deep feature learning for pedestrian gender recognition, Journal of the Franklin Institute (2017), http://dx.doi.org/10.1016/j.jfranklin.2017.09.003