Pedestrian Attribute Recognition using Gabor Wavelet Layers

Imran N. Junejo

Zayed University, Dubai, 19282, U.A.E.

Abstract. We address the problem of Pedestrian Attribute Recognition (PAR) in this paper. Owing to the presence of surveillance cameras in almost all outdoor and indoor public spaces, keeping and eye on pedestrian is a sought-after task with many useful applications. The problem entails recognizing attributes such as age-group, clothing style, accessories, footwear style etc. This is a multi-label problem and challenging even for human observers. We propose using a convolution neural network (CNN) with trainable Gabor wavelets (TGW) layers. The proposed layers are learnable and adapt to the dataset for a better recognition. The proposed multi-branch neural network is a mix of TGW and convolutional layers and we show its effectiveness on a public dataset.

Keywords: Gabor Wavelets, Convolutional Neural Networks, Pedestrian Attributes.

1 Introduction

Pedestrian attribute recognition is one of the active areas of research in the field of computer vision. The pedestrian attribute recognition deals with identifying a number of visual attributes from an image data. The identified attributes can belong to different classes, e.g. clothing style, footwear, gender, age group etc. A successful outcome of this research can be applied to various domains. It can be employed for motion analysis [20], where it can be used to identify crowd behavior attributes. Another important area of application is image-based surveillance or visual features extractions for person identification [18, 19]. Other applications include video analytics for business intelligence, or searching a criminal database for suspects using the identified visual attributes. Various factors make this a challenging problem. One of the main factors that makes this problem very difficult is the varying lighting conditions. Attributes of the same type of clothing can appear completely different under different lighting conditions. For example, distinguishing between black and dark blue colors is very difficult in certain weather conditions. Both colors will appear very similar to the camera in a darker environment. Occlusion also complicates the correct visual attribution identification and recognition. Occlusions can be either complete or partial and can results due to the camera orientation or from object self occlusions. For example, if a person is wears a hat, it might appear partially in the image, or its shape might be completely different. Similarly, the orientation of a person or a camera can hide a backpack partially or completely from the view. These examples clearly show that settings of an acquisition environment for image or video capture result in a high intra-class variations for the same visual attributes.

David C. Wyld et al. (Eds): CMLA, CIoT, CSEIT, NCS, NLPD, SPM - 2021
pp. 09-19, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111502

The focus of this work is the identification of visual attributes from image and video data. The distance of an object from the camera affects how that object appears in the image. If an object is very far from the camera, or if the image resolution is very low, a visual attribute, e.g. dress, hat, backpack, scarf, shoes etc. will only occupy a few pixels in the image. The combination of low image resolution, in addition to the self-occlusions or view-oriented occlusions, makes visual attribute identification a very challenging problem. Many of these issues can be seen in the most widely used pedestrian dataset. Figure 1 shows some of the samples from the PEdesTrian Attribute (PETA) [8]. PETA is the largest benchmark dataset. It comprises of 19000 images of different resolution that cover more than 60 attributes. The dataset is acquired from real-world surveillance camera systems and includes images of $8,705$ persons. It is a very challenging dataset because of the acquisition setup and scene settings. As can be seen in Figure 1, the quality of images is very low as well. This is due to a number of factors: images are very low resolution, acquisition problems result in a significant blur, many of the attributes are hidden due to severe occlusions. Moreover, due to the fast motion or acquisition problems some of the objects appear quite blurred thus making it a very challenging problem.

Visual attribute recognition problem can be solved in different ways, but the predominant solutions involve a two step process. In the first step, a feature extraction algorithm is employed to find a feature representation of the attributes. A number of feature extraction solutions are discussed in the computer vision literature. Most of these techniques require a very expert domain knowledge, and also needs a very high level of fine tuning for an accurate representation of visual attributes. For feature representation, methods like SIFT [16], HoG [7] or Haar-like features [25] have been employed in the field rigorously. Feature extraction is followed by the attributes classification step. For classification, Support Vector Machines (SVM) [8] has been the most widely used technique in the last decade.

In recent years, the convolutional neural networks (CNNs) have almost completely replaced SVMs for classification tasks. Compared to earlier attribute learning or image classification methods, CNNs are more effective and robust. Sarfraz et al. [23] proposed an end-to-end CNN-based network (VeSPA). This network had four parts, where each part corresponds to a specific pose category. Pose-specific attributes of each category are learned by each of these network parts. Their work demonstrated that coarse body pose information greatly influences the pedestrian attribute recognition. They extended their work in [21] and added a ternary view classifier in a modified approach that employed a global weighting solution. In this work, the global weighting solution for feature maps was employed before the final embedding. P-Net [2] employs a part-based approach. Based on GoogLeNet, the method guides the refined convolutional feature maps to capture different location information for the attributes related to different body parts. A joint person re-identification and attribute recognition approach (HydraPlus-Net) is presented by Liu et al. [15]. HydraPlus-Net is an Inception-based network and aggregates feature layers from multi-directional attention modules for the final feature representation. Sarafianos et al. [22] presented a multi-branch network that employed a simple weight scheme to address the class imbalance problem. They extracted visual attention masks to guide the network to crucial body parts. The masks are then fused at

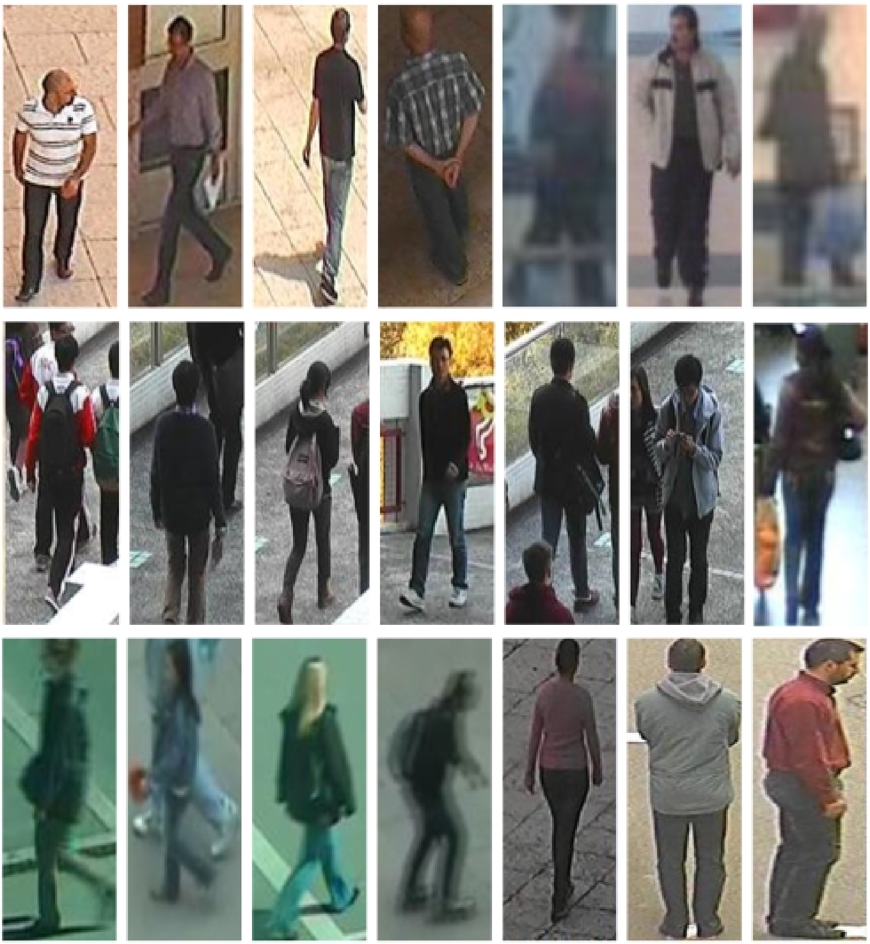

Fig. 1: PETA [8] dataset Samples.

different scales to obtain a better feature representation. Another end-to-end method for person attribute recognition that uses Class Activation Map (CAM) network [27] to refine attention heat map is proposed by Guo et al [10]. The heat map identifies the areas of different image attributes. They use CAM network to refine the attention heat map for an improved recognition. A Harmonious Attention CNN (HA-CNN) based joint learning approach for person re-identification is presented in [14]. They used HA-CNN for the joint learning of hard regional attention and soft pixel attention. Feature representation is obtained by this simultaneous optimization. A Multi-Level Factorization Net (MLFN) that factors the visual appearance of a person into latent discriminative factors is proposed by [4]. The factorization is done without manual annotation at multiple semantic levels. A Transferable Joint Attribute-Identity Deep Learning (TJ-AIDL) model that allows for a simultaneous learning of an identity discriminative and

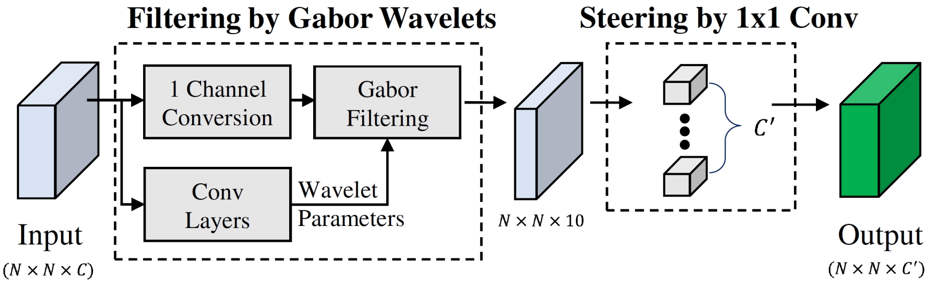

Fig. 2: Trainable Gabor Wavelet (TGW) layer [11]: Inputs and outputs are multichannel. A neural network is used to generate Gabor wavelet hyperparameters. These generated Gabor filters are then applied to the input. 1×1 convolution layer is added to enable the steerability of the Gabor wavelets.

attribute-semantic feature representation is proposed by [26]. Si et al. [24] proposed a Dual ATtention Matching network (DuATM), which is a joint learning end-to-end person re-identification framework. Their method simultaneously performs context-aware feature sequences learning and attentive sequence comparison in a joint learning mechanism for person re-identification.

Using Gabor wavelets with CNNs have received a tremendous attention as well [1, 3, 11, 17]. [1] use a Gabor filter bank as the first layer of a CNN and the bank gets updated using the standard back-propagation network leaning phase. [3] also use Gabor filters in the first layer of the network. While introducing lateral inhibition to enhance network performance, they use a n-fold cross validation to search for the best parameters. Authors in [17] introduce a Gabor Neural Network (GNN) where Gabor filters are incorporated into the convolution filter as a modulation process, in a spirit similar to the above mentioned works. In contrast to the above works where fixed Gabor filters are used, [11] introduce a trainable Gabor wavelets (TGW) layer. The authors present a method where the hyperparameters of the wavelets are learned from the input and a novel 1×1 convolution layers are employed to create steerable filters. In this paper, we propose using this TGW layer with our proposed CNN for a novel solution to the problem of PAR. We test on a challenging dataset and show a considerable improvement over state of the art.

2 Main Approach

In this section, we start with the description of the Gabor wavelet layer. Then we describe the architecture of our network in general.

2.1 Gabor Wavelet Layer

We make use of the Trainable Gabor wavelets (TGW) layer as proposed by Kwon et. al. [11] (see. Fig. 2). A neural network is used to generate the hyperparameters for the

Gabor wavelet and the generated Gabor filters are applied to filter inputs. In order to capture essential input features, a 1×1 convolution layer is added to the TGW layer to capture features at different orientations.

Hyperparameter estimation The 2D Gabor wavelet can be described as:

$$G(x, y) = \exp\left(-\frac{X^2 + \gamma Y^2}{2\sigma^2}\right) \times \cos\left(\frac{2\pi}{\lambda} X\right) \tag{1}$$

where γ represents aspect ratio, λ represents wavelength of the sinusoidal, σ represents width or the standard deviation, $X = x\cos(\theta) + y\sin(\theta)$, $Y = -x\sin(\theta) + y\cos(\theta)$, and θ is an angle in the range $[0, \pi]$. Thus in order to specify a continuous Gabor wavelet, we need to determine the set of hyperparameters $\{\gamma, \theta, \lambda, \sigma\}$. In order to convert the continuous filter to a discrete one, a sampling grids need to be defined, which is largely linked to σ. A new parameter is thus introduced to compute the discrete filter:

$$G[m, n] = g(u, v) = \left(\frac{m}{\lfloor\zeta\rfloor} \times \zeta, \frac{n}{\lfloor\zeta\rfloor} \times \zeta\right) \tag{2}$$

where m and n are in the interval $-\lfloor\zeta\rfloor, \lfloor\zeta\rfloor + 1, \ldots, \lfloor\zeta\rfloor$, and by just varying $\lfloor\zeta\rfloor$, variety of sampling grids can be achieved [11]. For a loss function L, we need to compute $\frac{\partial L}{\partial \zeta}$ in order to train for the wavelet layer that is cascaded with our CNN. In order to train for the ζ, what remains is to compute $\frac{\partial G[m,n]}{\partial \zeta}$, as $\frac{\partial L}{\partial G[m,n]}$ is handled automatically by the deep learning libraries:

$$\begin{aligned} \frac{\partial G[m, n]}{\partial \zeta} &= \frac{\delta g(u, v)}{\partial u} \frac{\partial u}{\partial \zeta} + \frac{\partial g(u, v)}{\partial v} \frac{\partial v}{\partial \zeta} & (3) \\ &= \frac{\delta g(u, v)}{\partial u} \frac{u}{\zeta} + \frac{\partial g(u, v)}{\partial v} \frac{v}{\zeta} & (4) \end{aligned}$$

as $\frac{d}{d\zeta}\lfloor\zeta\rfloor = 0$. The remaining parameters $\frac{\partial G[m,n]}{\partial \sigma}$, $\frac{\partial G[m,n]}{\partial \gamma}$, $\frac{\partial G[m,n]}{\partial \lambda}$ can be computed in a similar way and a similar parameterization can be adopted for the parameters σ, γ and λ.

A very significant parameter for the Gabor wavelet is the orientation (θ). These values are mostly chosen empirically. This parameter is also made trainable to better design orientations for the task at hand. To use the steering property, where a linear combination of finite set of responses can be used to represent convolution at any orientation, a 1×1 convolution layer, working as a linear combination layer, is added to the output of the generated filters. For this layer, ten equally spaced fixed orientations are selected, working as basis filters: $9°, 27°, 45°, 63°, 81°, 99°, 117°, 135°, 153°$, and $171°$ [11].

2.2 Attribute Recognition Network

The above mentioned TGW layer can be thought of as a feature extracting layer. In addition to this, we also employ it as the key building block of our network. Thus, in

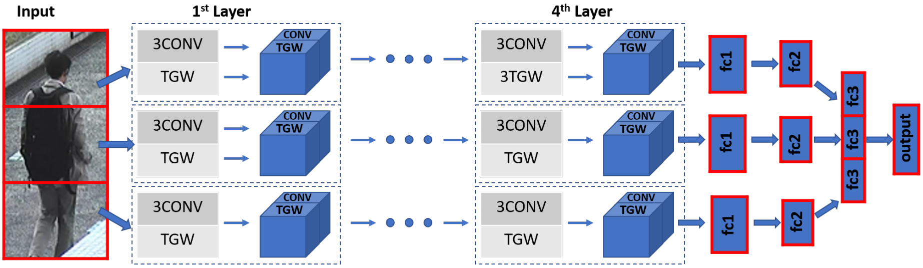

Fig. 3: Our Approach: The proposed method divides the input image into three parts. For each branch, the network contains 4 layers that are a mix between TGW and 3Conv layer (mixed-layers). The output of each branch is followed by three fc layers. Size of the last layer of the network matches the number of attributes of the dataset. Parameters of the network are mentioned in Table 1.

addition to functioning as the *lowest layer*, it also aids the network to learn high level features.

The proposed network is shown in Fig. 3. An input image is divided into three equal parts along on the vertical axis. Each part of the image passes through a separate branch of the network. As can be seen in the figure, each branch consists of 4 mixed-layers: combination of TGW layer and a 3×3 convolution layer. The input to the TGW layer starts with a 1-channel conversion, i.e. a multi-channel input is converted to a 1-channel, which is a summation over the channels operation for all layers except the first layer where we perform a simple color-to-gray image conversion. The parameters for these layers are given in Table 1.

Each mixed-layer (1 to 4) contains 256 channels from the TGW layer and 256 channels from a $3{\times}3$ convolution layer (denoted as 3Conv). Thus depth of each mixed-layer output is 512 (concatenation of TGW and 3Conv layer). The network thus contains blocks of layers stacked together. For each 3Conv layer, as the name suggest, the kernel size is 3×3. The convolution is followed by ReLU activation function, max-pool layer (size 2×2), and Batch Normalization (BN) layer. The size of the input image to each of these stacked layers is, respectively: 48×48, 24×24, 12×12, and 6×6.

Output from each branch encounters three fully connected layers, i.e. fc1, fc2 and fc3, of size 512, 512 and 35, respectively. Each fc layer uses ReLU as the activation function, followed by a dropout layer ($p = 0.5$), to minimize the number of parameters of the network. fc3 from all branches are concatenated and the final output layer size matches the number of dataset attributes.

The method proposes using Gabor wavelets embedded with a deep neural network. Whereas other methods construct Gabor filters manually, the proposed network learns the wavelet parameters suitable to the dataset. Generated Gabor filters are stacked with convolution layers to build the overall network. As we shall show next, the proposed network is efficient and learns the dataset structure well to perform at par with state of the art.

Layer	γ_o	λ_o	σ_o	ζ_o	TGW Channels	Conv Channels
1	0.3	6.8	5.4	6	256	256
2	0.3	5.6	4.5	5	256	256
3	0.3	4.6	3.6	4	256	256
4	0.3	3.5	2.8	3	256	256

Table 1: Parameters used for the TGW layers.

3 Evaluation

Following channel conversion, the grayscale image is divided into three parts. Each part of the networks encounters 4 mixed-layers, consisting of equal number of channels from TGW and 3`Conv` layer. Depth of each mixed-layer is 512. The mixed-layers are followed by a series of fully connected layers before the final output layer. `ReLU` is used as the activation function for all the layer. The output layer uses `sigmoid` as the activation function.

In order to evaluate our method quantitatively, we compute various measures and report the results below. Although mean accuracy has been widely used in the attribute recognition literature, it treats each attribute independent of the other attributes. This might not necessarily be the case and an inter-attribute correlation might exist. Therefore, researchers also report *example-based* evaluations, namely accuracy (Acc), precision ($Prec$), recall (Rec), and F1 score ($F1$) [13].

3.1 Dataset

PETA is one of the most widely used dataset for the problem of pattern attribute recognition. Collected from real-time surveillance cameras, the PETA dataset contains $19,000$ images collected from 10 publicly available datasets. The resolution of the images ranges from 17×39 to 169×365. Most of the previous works [12, 23] report results on the PETA dataset using only 35 attributes. Similarly, for a fair comparison, experiments are conducted on 5 random splits: we allocate $9,500$ samples for training, $1,900$ samples for validation, $7,600$ samples for testing on the dataset.

Pre-processing: Before continuing to the next step, we perform **mean subtraction**: That is, we compute the mean for all the images for each color spaces and this value is subtracted from image data. Intuitively for each dimension, this step is equal to centering the data around its origin. Next step involves **normalization**: We compute the standard deviation separately for each color space and the image data is divided by this value.

3.2 Setup

For deep learning, we adopted the KERAS [6] library, which is based on the TensorFlow backend. All experiments were performed on a cluster node with 2 x Intel Xeon E5 CPU, 128GB Registered ECC DDR4 RAM, 32TB SAS Hard drive storage, and 8 x NVIDIA Tesla K80 GPUs.

	PETA [8]			
	Acc	$Prec$	Rec	$F1$
Chen et. al. [12]	75.07	83.68	83.14	83.41
Liu et. al. [28]	74.62	82.66	85.16	83.40
Sarfaraz et. al. [23]	77.73	86.18	84.81	85.49
ours	$\mathbf{79.35}$	$\mathbf{86.24}$	79.45	81.48

Table 2: Quantitative results (%) on PETA datasets. Results are compared with the other benchmark methods. As can be seen, we have comparable results, with considerable improved accuracy for the datasets.

3.3 Implementation Details

We train the network for 50 epochs. `ReLU` was used as the activation function for all layers of the network. We used the `Adam` for update optimizer using the parameters: learning rate = $1e^{-4}$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

We added the dropout layers to the fc layers to prevent model over-fitting. We adopt weight decay by a factor of 0.1 after 15 epochs. The batch size was set to be 8. All weights in the network are initialized using `He Normal` initialization.

For the TGW layers with a steering block, we use the scheme suggested by [9]: we fix the parameters $\{\gamma, \sigma, \lambda\}$ as shown in Table 1 while training for ζ. This setup yields the best results in our experiments.

3.4 Results

We evaluate the effectiveness of the proposed method on PETA datasets. Table 2 shows a comparison of the proposed method with six current state of the art methods. For the PETA dataset, Acc obtained from our method is 79.35%. This is higher than all the other methods that we compare with. The obtained results for the other measures (Pre, Rec and $F1$) is 86.24%, 79.45%, and 81.48% respectively. Class-wise accuracy chart for the PETA dataset is shown in Fig. 4. Interestingly, the lowest accuracy is that for the class `upperBodyOther`. Considering the image resolutions in the dataset, this is indeed a very difficult class to accurately measure. On the other hand, the highest accuracy is that of the classes `upperBodyThinStripes` and `upperBodyVNeck`.

The proposed method makes a novel use of the Gabor wavelet layers. Instead of manually constructing Gabor filters, the layers are trainable and are able to correctly estimate model parameters. The method divides input image into three parts. For each part, we train four mixed-layers: combination of TGW and `3Conv` layers. The output of these branches are concatenated and then followed by three fc layers. We have obtained very encouraging results for the key measures. The method is novel and unique in the sense that it does not resort to data augmentation or part-based computations, as employed by [13]. We also do not have to compute pose estimation [12], or construct any hand-crafted features [5]. Our results are an improvement over state of the art and clearly justifies the use of Gabor wavelet layers.

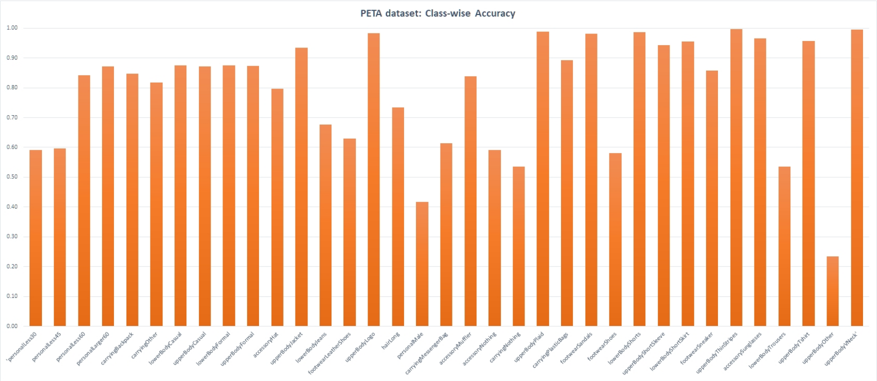

Fig. 4: Class-wise Accuracy - PETA dataset: the figure shows the obtained class-wise accuracy. The highest accuracy is for the class `upperBodyThinStripes,upperBodyVNeck`. The lowest accuracy is 23.4% for the class `upperBodyOther`.

4 Conclusion

This work proposes the idea of using trainable Gabor wavelets (TGW) for the task of pedestrian attribute recognition. We have proposed a multi-branch neural network. The input to the network is an image that is divided into three parts, each processed through a different branch of the network. Each branch contains mixed-layers that are capable of learning the Gabor wavelet parameters. The filters in each branch are learned from the data itself. We have tested the data on a challenging public dataset and are encouraged by the results. In future work, we aim to experiemnt with other publicly available datasets with possibly different network architectures.

References

1. Alekseev, A., Bobe, A.: Gabornet: Gabor filters with learnable parameters in deep convolutional neural network. In: 2019 International Conference on Engineering and Telecommunication (EnT). pp. 1–4 (2019). https://doi.org/10.1109/EnT47717.2019.9030571
2. An, H., Fan, H., Deng, K., Hu, H.M.: Part-guided network for pedestrian attribute recognition. 2019 IEEE Visual Communications and Image Processing (VCIP) pp. 1–4 (2019)
3. Bai, J., Zeng, Y., Zhao, Y., Zhao, F.: Training a v1 like layer using gabor filters in convolutional neural networks. In: 2019 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2019)
4. Chang, X., Hospedales, T.M., Xiang, T.: Multi-level factorisation net for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
5. Chen, Y., Duffner, S., STOIAN, A., Dufour, J.Y., Baskurt, A.: Pedestrian attribute recognition with part-based CNN and combined feature representations. In: Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. pp. 114–122 (2018)

6. Chollet, F.: keras (2015), https://github.com/fchollet/keras
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 1, pp. 886–893 (2005)
8. DENG, Y., Luo, P., Loy, C.C., Tang, X.: Pedestrian attribute recognition at far distance. In: Proceedings of the 22nd ACM International Conference on Multimedia. pp. 789–792. MM '14 (2014)
9. Guo, G., Mu, G., Fu, Y., Huang, T.: Human age estimation using bio-inspired features. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009. pp. 112 – 119 (07 2009). https://doi.org/10.1109/CVPR.2009.5206681
10. Guo, H., Fan, X., Wang, S.: Human attribute recognition by refining attention heat map. Pattern Recognition Letters **94**(C), 38–45 (Jul 2017)
11. Kwon, H.J., Koo, H., Soh, J.W., Cho, N.I.: Age estimation using trainable gabor wavelet layers in a convolutional neural network. 2019 IEEE International Conference on Image Processing (ICIP) pp. 3626–3630 (2019)
12. Li, D., Chen, X., Zhang, Z., Huang, K.: Pose guided deep model for pedestrian attribute recognition in surveillance scenarios. In: 2018 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6 (2018)
13. Li, D., Zhang, Z., Chen, X., Ling, H., Huang, K.: A richly annotated dataset for pedestrian attribute recognition. CoRR **abs/1603.07054** (2016)
14. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
15. Liu, X., Zhao, H., Tian, M., Sheng, L., Shao, J., Yan, J., Wang, X.: Hydraplus-net: Attentive deep features for pedestrian analysis. In: Proceedings of the IEEE international conference on computer vision. pp. 1–9 (2017)
16. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the Seventh IEEE International Conference on Computer Vision. vol. 2, pp. 1150–1157 (1999)
17. Luan, S., Zhang, B., Zhou, S., Chen, C., Han, J., Yang, W., Liu, J.: Gabor convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1254–1262 (2018). https://doi.org/10.1109/WACV.2018.00142
18. Nanda, A., Chauhan, D.S., K. Sa, P., Bakshi, S.: Illumination and scale invariant relevant visual features with hypergraph-based learning for multi-shot person re-identification. Multimedia Tools and Applications **78**(4), 3885–3910 (Feb 2019)
19. Rahman, K., Abdul Ghani, N., Abdulbasah Kamil, A., Mustafa, A., Kabir Chowdhury, M.A.: Modelling pedestrian travel time and the design of facilities: A queuing approach. PLOS ONE **8**(5), 1–11 (05 2013). https://doi.org/10.1371/journal.pone.0063503
20. Raudies, F., Neumann, H.: A bio-inspired, motion-based analysis of crowd behavior attributes relevance to motion transparency, velocity gradients, and motion patterns. PLOS ONE **7**(12), 1–17 (12 2013). https://doi.org/10.1371/journal.pone.0053456
21. Saquib Sarfraz, M., Schumann, A., Eberle, A., Stiefelhagen, R.: A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
22. Sarafianos, N., Xu, X., Kakadiaris, I.A.: Deep imbalanced attribute classification using visual attention aggregation. In: Springer European Conference on Computer Vision. pp. 708–725 (2018)
23. Sarfraz, M., Schumann, A., Wang, Y., Stiefelhagen, R.: Deep view-sensitive pedestrian attribute inference in an end-to-end model. In: Bristish Machine Vision Conference (BMVC) (09 2017)

24. Si, J., Zhang, H., Li, C.G., Kuen, J., Kong, X., Kot, A.C., Wang, G.: Dual attention matching network for context-aware feature sequence based person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
25. Viola, P., Jones, M.: Robust real-time object detection. In: International Journal of Computer Vision (IJCV). vol. 57 (01 2001)
26. Wang, J., Zhu, X., Gong, S., Li, W.: Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
27. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1. pp. 487–495. NIPS'14, MIT Press, Cambridge, MA, USA (2014)
28. Zhou, Y., Yu, K., Leng, B., Zhang, Z., Li, D., Huang, K.: Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization. In: British Machine Vision Conference BMVC 4-7 (2017)