6-30-2021

Pedestrian attribute recognition using trainable Gabor wavelets

Imran N Junejo
Zayed University

Naveed Ahmed
University of Sharjah

Mohammad Lataifeh
University of Sharjah

Heliyon 7 (2021) e07422

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Pedestrian attribute recognition using trainable Gabor wavelets

Imran N. Junejo [a,*], Naveed Ahmed [b], Mohammad Lataifeh [b]

[a] Zayed University, Dubai, United Arab Emirates
[b] University of Sharjah, Sharjah, United Arab Emirates

ARTICLE INFO

Keywords:
Deep learning
Attribute recognition
Computer vision

ABSTRACT

Surveillance cameras are everywhere keeping an eye on pedestrians or people as they navigate through the scene. Within this context, our paper addresses the problem of pedestrian attribute recognition (PAR). This problem entails the extraction of different attributes such as age-group, clothing style, accessories, footwear style etc. This is a multi-label problem with a host of challenges even for human observers. As such, the topic has rightly attracted attention recently. In this work, we integrate trainable Gabor wavelet (TGW) layers inside a convolution neural network (CNN). Whereas other researchers have used fixed Gabor filters with the CNN, the proposed layers are learnable and adapt to the dataset for a better recognition. We test our method on publicly available challenging datasets and demonstrate considerable improvements over state of the art approaches.

Introduction

Being one of the active areas of research in computer vision, the pedestrian attribute recognition (PRA) deals with identifying several visual attributes from an image data. The identified attributes 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 [1, 2], 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 and tracking [3, 4, 5], all of which can lead to further applications such as video analytic for business intelligence, and person re-identification based on the extracted features [6].

Various factors add to the complexity of this challenge. One of the main factors is the changing lighting conditions. Attributes of the same type of clothing or objects can appear completely different under various 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 [7]. Complete or partial occlusions occur due to the camera orientation, or from object self-occlusions. For example, if a person is wearing 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 variation for the same visual attributes.

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 an image. If the object is very far from the camera or if the image resolution is very low, a visual attribute for a 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 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 datasets. Fig. 1 shows some of the samples from the PEdesTrian Attribute (PETA) [8] and A Richly Annotated Pedestrian (RAP) [9] datasets. PETA is one of the largest benchmark datasets. 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 Fig. 1, the quality of images is very low as well. This is mainly due camera resolution, acquisition conditions producing significant blur, and few occlusions that cause many attributes to remain hidden. RAP dataset comprises of 41 thousand images covering 72 attributes and is acquired from multiple viewpoints. The dataset shows a huge variation in the attributes due to pedestrian appearances, viewpoints, and severe occlusions. An analysis of these datasets reveals that visual attributes identification from these images is an extremely diffi-

* Corresponding author.
E-mail address: imran.junejo@zu.ac.ae (I.N. Junejo).

https://doi.org/10.1016/j.heliyon.2021.e07422
Received 1 February 2021; Received in revised form 4 April 2021; Accepted 24 June 2021

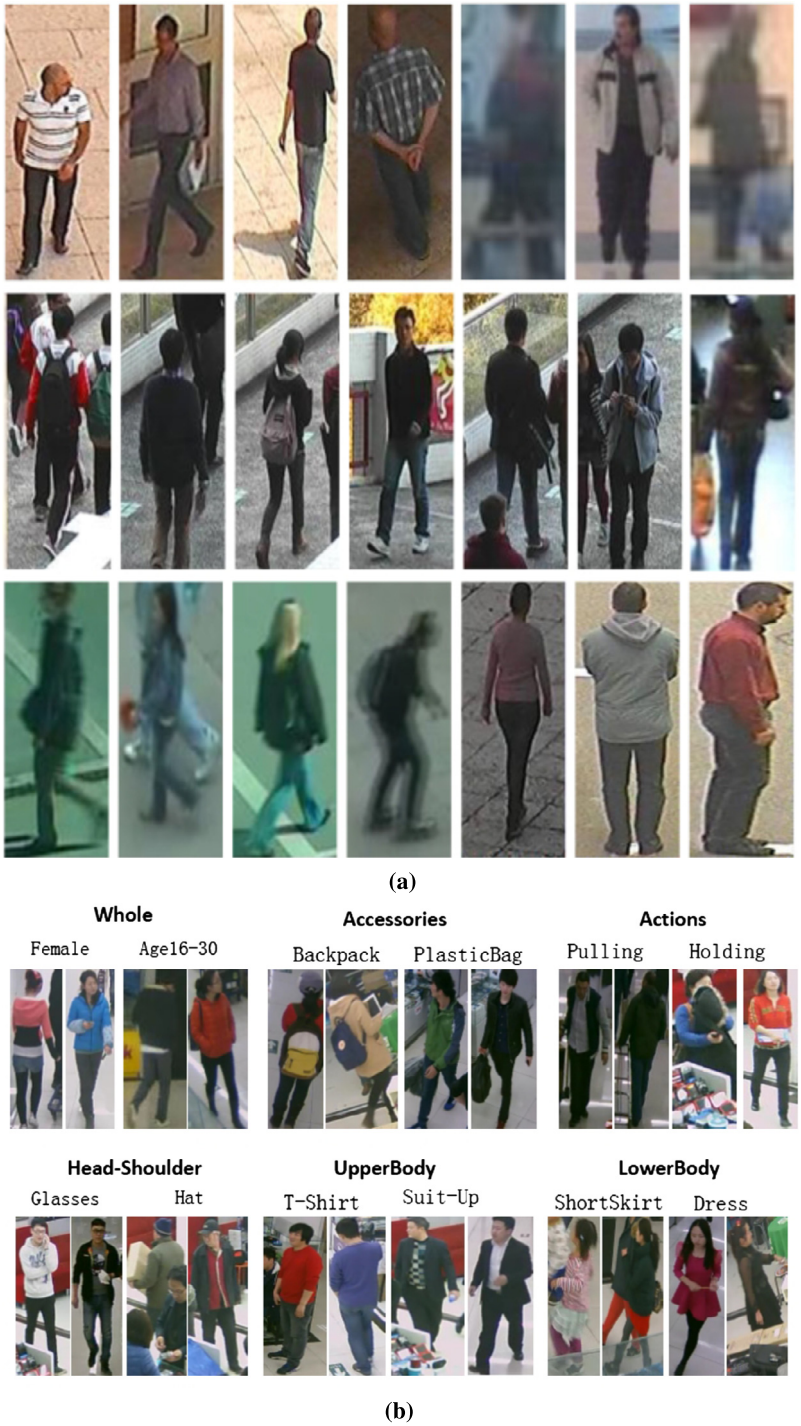

Fig. 1. (a) PETA [8] dataset Samples. (b) RAP [9] dataset samples.

cult task due to the very low image quality. Many of the attributes are largely occluded as well. Moreover, some of the objects appear quite blurred due to the fast motion or acquisition problems - adding more complexity to the 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. Different feature extraction solutions are discussed in the computer vision literature. Most of these techniques require deep domain-knowledge and high-level expertise in fine tuning for an accurate representation of visual attributes. For feature representation, methods like SIFT [10], HoG [11] or Haar-like features [12] have been employed in the field rigorously. Feature extraction is followed by the attribute's classification step, for which, 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. In this work, we make use of the Gabor wavelets, which have been used in the computer vision literature extensively over the last many decades. However, there have been only few works that use the Gabor wavelets in conjunction with the CNNs. For the majority of the works that do employ these wavelets, filters are pre-

constructed and fed as filters to the convolutional network. However, we adopt an approach where the convolutional network is employed to learn the wavelet parameters along with learning other neural network parameters. These Trainable Gabor wavelets (TGW) [13] layers make up for the backbone of our network. Each TGW layer accepts a single channel input, with a multi-channel output, and learns the best parameters to generate an adaptable set of Gabor filters. TGW layer contains a 1×1 convolution layer that uses the steerability of Gabor wavelets to address orientation issues. We also use a regular convolutional layer to extract features from the input as well. These outputs from TGW and convolution layers are stacked together, which are referred to as *mixed-layer*, and make up the building block of our network. The proposed network, shown in Fig. 3, undergoes a series of fully connected (fc) layers that are connected to the final network output layer. The proposed network is simple and trainable with a standard gradient-decent method.

Our main contributions are:

- We for the first time make use of the trainable Gabor wavelets to the problem of pedestrian attribute recognition.
- We propose a novel network that, while learning the Gabor wavelet parameters, combines the learned wavelet features with the regular convolution layers.
- The proposed method is demonstrated to have better recognition results than state of the art on two of the most challenging public datasets.

Related work

In this section we discuss works that are in spirit similar to our method, a detailed survey can be found in Wang et al. recent survey [14]. PETA [8] is one of the most widely used pedestrian datasets. While introducing the dataset, the Deng et al. [8] use the luminance channel and apply Ensemble of Localized Features (ELF) and Gabor and Schmid filter on it. To address the class imbalance problem, they also apply ikSVMs [15] separately on each attribute. They exploit similarity between images using the Markov Random Field (MRF). In their representation, each image is a node and the link between two nodes is determined by the similarity between neighboring images. RAP dataset [9] is acquired from multiple viewpoints that introduces significant variations for the same attributes along with severe occlusions. They employed two CNN models based on Caffe framework [16] to analyze the impact of the variations introduced by different viewpoints and occlusions on the overall classification of the attributes. They trained SVMs in addition to the adopting ELF. Additionally, they divide the image into multiple blocks (three in their case) to employ a part-based classification scheme. For their work, the parts comprised of: upper body (torso), lower body, and head and shoulders. Joo et al. [17] proposed another approach that also employed part-based recognition. In their work, they first created Histogram of Oriented Gradient (HoG) features from an image subdivided into multiple overlapping regions. For attributes classification, they employed a Poselet-based approach [18]. Furthermore, Zhao et al. [19] proposed a solution based on a Recurrent Neural Network (RNN). In their work, they employed two end-to-end models: Recurrent Attention (RA) and Recurrent Convolutional (RC). The correlations between various attribute groups are mined by the RC model, while the intra-group attention, correlation, and spatial locality are used by the RA model to improve the performance and robustness of pedestrian attribute recognition. Nonetheless, their network has a very deep architecture, hence the number of parameters is quite large. In another part-based approach, Zhu et al. [20] proposed a CNN-based solution where the human body is divided into 15 parts, and a CNN is trained separately for each part. The contribution of each attribute determines the weight of the corresponding CNN. Zhou et al. [21] use GoogLeNet for the initial mid-level feature extraction from detection layers. The activation maps from these detection layers are clustered and fused to localize the pedestrian attributes. Only image labels are used to train the detected layers in order to learn the relationship between the mid-level features and the pedestrian attributes. Max-pooling is used in a weakly-supervised technique for object detection training. Similarly, Chen et al. [22] suggested a part-based network that combined LOMO features [23] with CNN extracted features. They showed that the Scale-Invariant Local Ternary Patterns and HSV histograms based LOMO features are illumination-invariant texture and color descriptors.

Furthermore, a pose-guided model was also presented [24] based on pedestrian body structure knowledge. In the first step, the model computes transformation parameters to estimate the pose from an image. Based on the pose information, human body parts are localized, and the final attribute recognition is estimated by fusing multiple features. Another parts-localization method was offered by Liu et al. [25]. They proposed a Localization Guide Network (LGNet) that uses a CNN model based on Inception-v2 [26] for feature extraction. Afterwards, a global average pooling layer (GAP) is adopted to extract global features. The global and local features are fused to perform the pedestrian attributes classification. A visual semantic graph approach has also been presented [27], using ResNet-50 to for the pedestrian images feature extraction. Yet, having more than fifty layers, the proposed network contained a large number of parameters. Furthermore, a multi-branch approach has also been proposed using multi-colorspace input [28].

Sarfraz et al. [29] proposed an end-to-end CNN-based network (VeSPA). This network consists of 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. This work was extended in [30] adding a ternary view classifier in a modified approach as feature maps were obtained using a global weighting solution prior to the final embedding. P-Net [31] employs a part-based approach using GoogLeNet. The location attributes for different body parts are estimated using refined convolutional feature maps. A joint person re-identification and attribute recognition approach (HydraPlus-Net) is presented by Liu et al. [32]. HydraPlus-Net is an Inception-based network and aggregates feature layers from multi-directional attention modules for the final feature representation. Sarafianos et al. [33] present a multi-branch network that addresses class imbalance problem by employing a trivial weighting scheme. The network is guided towards crucial body parts using the extracted visual attention masks. These visual attention masks are used to obtain an improved feature representation by fusing them at varying scales. Another end-to-end method for person attribute recognition that uses Class Activation Map (CAM) network [34] to refine attention, heat map is proposed by Guo et al. [35] where different image attributes are identified using 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 [36]. Hard regional and soft pixel attention are learned in a combined manner using HA-CNN. Feature representation is obtained by this simultaneous optimization. A Multi-Level Factorization Net (MLFN) that identifies latent discriminative factors from visual appearance of a person is proposed by [37]. The multi-semantic levels factorization is done without manual annotation. A Transferable Joint Attribute-Identity Deep Learning (TJ-AIDL) model that allows for a simultaneous learning of an identity discrimination and attribute-semantic feature representation is proposed by [38]. Furthermore, Si et al. [39] 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.

A Generative Adversarial Network based pose-normalized person re-identification framework is presented in [40]. They learn pose invariant deep person re-identification features using synthesized images. A deep

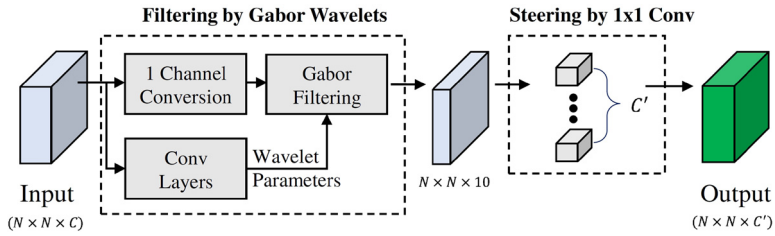

Fig. 2. Trainable Gabor Wavelet (TGW) layer [13]: Inputs and outputs are multichannel. A neural network is used to generate Gabor wavelet hyperparameters. These generated Gabor filters are then applied to the input. 1×1 convolution layer is added to enable the steerability of the Gabor wavelets.

CNN based method to learn partial descriptive features for efficient person feature representation is presented in [41]. They employed a pyramid spatial pooling module and reported an improvement of 2.71% on the PETA dataset over [29]. [42] improved over [29] by employing a deeper network based on a context sensitive framework. Their proposed network creates a richer feature set using deeper residual (ResNet) networks that improved generalization and classification accuracy. The resulting model achieved the best in class results on attribute recognition datasets. Additionally, [43] presented a visual semantic graph reasoning framework that modeled spatial and attribute relationships using two types of graphs. A Graph Convolutional Network that combines potential semantic relationships of the attributes, and spatial relationship between local regions to be used for the training. A dual model approach was also presented for pedestrian recognition [44] using Recurrent Attention (RA) and Recurrent Convolutional (RC). The RC model employed a Convolutional-LSTM model to establish the correlations between different groups of attributes. To improve the overall robustness, the RA model uses both local attention correlation and global spatial locality.

Using Gabor wavelets with CNNs have received a tremendous attention as well [13, 45, 46, 47, 48]. Gabor filter bank was proposed as the first layer of a CNN [45], the bank gets updated using standard back-propagation network learning phase. Similarly, Gabor filters were used in the first layer [46], however, while introducing lateral inhibition to enhance network performance, a n-fold cross validation was used to search for the best parameters. Within this approach, a combination of HOG and Gabor filters were used for feature extraction while CNNs were deployed for detection in [47]. Authors in [48] introduce a Gabor Neural Network (GNN) where Gabor filters are incorporated into the convolution filter as a modulation process, in spirit similar to the above mentioned works. In contrast to the above works where fixed Gabor filters are used, [13] introduce a trainable Gabor wavelet (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. Our method is tested on two of the most challenging datasets and shows a considerable improvement over state of the art approaches.

1. Main approach

In this section, we start with the description of the Gabor wavelet layer. Followed by the proposed architecture of the network.

1.1. Gabor wavelet layer

We make use of the Trainable Gabor wavelets (TGW) layer as proposed by Kwon et al. [13] (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.

1.1.1. 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, 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 [13]. 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:

$$\frac{\partial G[m, n]}{\partial \zeta} = \frac{\delta g(u, v)}{\partial u}\frac{\partial u}{\partial \zeta} + \frac{\partial g(u, v)}{\partial v}\frac{\partial v}{\partial \zeta} \tag{3}$$

$$= \frac{\delta g(u, v)}{\partial u}\frac{u}{\zeta} + \frac{\partial g(u, v)}{\partial v}\frac{v}{\zeta} \tag{4}$$

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 $171°$ [13].

1.2. 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 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 first converted to a grayscale and then passes through a series of 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 channel's 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.

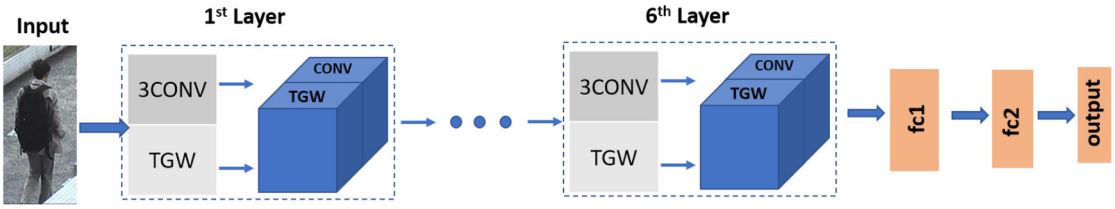

Fig. 3. Our Approach: The input images go through a series of 6 mixed-layers. The output of layer six is followed by three fc layers. Size of the last layer of the network matches the number of dataset attributes. Parameters of the network are mentioned in Table 1.

Table 1. Parameters used for the TGW layers.

Layer	γ_o	λ_o	σ_o	ζ_o	TGW Channels	Conv Channels
1	0.3	6.8	5.4	6	128	128
2	0.3	5.6	4.5	5	128	128
3	0.3	4.6	3.6	4	128	128
4	0.3	3.5	2.8	3	128	128
5	0.3	2.5	2.0	2	128	128
6	0.3	2.5	2.0	2	128	128

Each mixed-layer (1 to 6) contains 128 channels from the TGW layer and 128 channels from a 3×3 convolution layer (denoted as `3Conv`). Thus, depth of each mixed-layer block output is 256 (concatenation of TGW and `3Conv` layer). For each `3Conv` layer, as the name suggest, the kernel size is 3×3. The convolution is followed by `LeakyReLU` activation function, max-pool layer (size 2×2), and a Batch Normalization (BN) layer. The size of an input image to each of these stacked layers is, respectively: 227×227, 113×113, 56×56, 28×28, 14×14, and 7×7.

The mixed-layers are followed by three fully connected layers, i.e., $fc1$, $fc2$ and output, of size 512, 512 and 35 for PETA or 51 for RAP, respectively. Each fc layer uses `LeakyReLU`(0.01) as the activation function, followed by a dropout layer ($p = 0.5$), to minimize the number of parameters in the network. The final layer size matches the number of attributes of the dataset.

The method proposes using Gabor wavelets merged with a deep neural network. Whereas other methods construct Gabor filters manually, 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 structure of the dataset well to perform at par with state of the art.

2. Evaluation

As mentioned above, following the channel conversion, the grayscale image is processed through mixed-layers. Each mixed-layer consists of equal number of channels from TGW and `3Conv` layer. Depth of each mixed layer output is 256. The mixed-layers are followed by a series of fully connected layers before the final output layer. `LeakyReLU`(0.01) is used as the activation function for all the layers. The output layer uses `sigmoid` as the activation function.

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 independently 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 ($F1$) [9].

2.1. Dataset

RAP and PETA are the most widely used datasets 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. Collected from a multi-camera setup of around 26 cameras, the RAP dataset contains 41,585 pedestrian samples. Each attribute is annotated independently, and the size of the images range from 36×92 to 344×554.

Most of the previous works [24, 29] report results on the PETA dataset using only 35 attributes. Similarly, for the RAP dataset, results are reported on 51 datasets. In order to make a fair comparison, we adopt the same scheme and test/train on the same 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 PETA dataset. For the RAP dataset, we split it randomly into 33,268 training images and 8,317 test images [29]. We adopted the weighted-cross entropy loss function [24] in order to mitigate the class imbalance problem. Similarly, following other researchers, images are resized to an image resolution of 144×48.

Pre-processing: we start with what is known as the *mean subtraction* where mean is computed for all images (for each of the color channel) and subtracted from the image data. Similarly, we compute the standard deviations, the *normalization step*, for images (and their color channels) and divide image values by this statistic. These steps are crucial and are equivalent to centering the data around its origin.

2.2. Setup

For deep learning, we adopted the KERAS [49] library, which is based on the TensorFlow backend. All experiments were performed on a cluster node with 2 x Intel Xeon E5 CPU, 128 GB Registered ECC DDR4 RAM, 32TB SAS Hard drive storage, and 8 x NVIDIA Tesla K80 GPUs.

2.3. Implementation details

We train the network for 50 epochs. `LeakyReLU` was used as the activation function for all layers of the network with the parameter 0.01. 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 overfitting. 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 [50]: 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.

2.4. Results

We evaluate the effectiveness of the proposed method on both PETA and RAP 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 80.04%. This is higher than all the other methods that we compare with. The obtained results for the other measures (Pre, Rec and $F1$) is 86.49%, 80.1%, and 82.32% 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

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

	PETA [8]				RAP [9]			
	Acc	*Prec*	*Rec*	*F*1	*Acc*	*Prec*	*Rec*	*F*1
Chen et al. [24]	75.07	83.68	83.14	83.41	62.02	74.92	76.21	75.56
Li et al. [9]	–	–	–	–	63.67	76.53	77.47	77.00
Sudowe et al. [51]	73.66	84.06	81.26	82.64	62.61	80.12	72.26	75.98
Liu et al. [21]	74.62	82.66	85.16	83.40	53.30	60.82	78.80	68.65
Sarfaraz et al. [29]	77.73	86.18	84.81	85.49	67.35	79.51	79.67	79.59
Li et al. [32]	76.13	84.92	83.24	84.07	65.39	77.33	78.79	78.05
ours	**80.04**	**86.49**	80.1	82.32	**91.1**	**92.39**	**91.1**	**91.56**

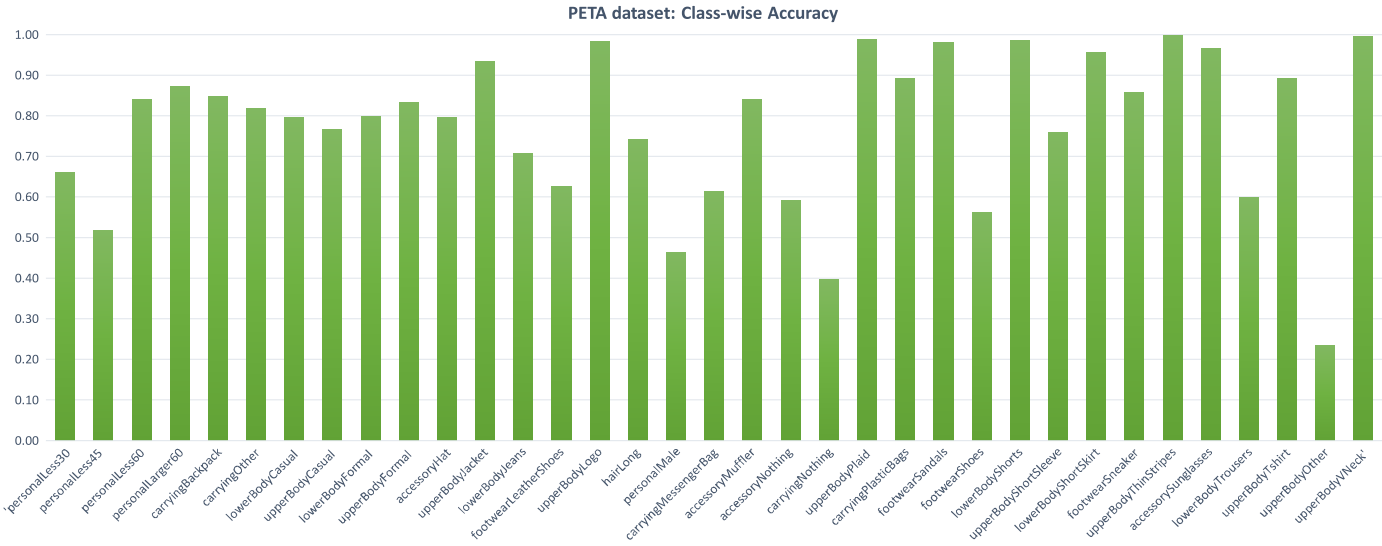

Fig. 4. Class-wise Accuracy - PETA dataset: the figure shows the obtained class-wise accuracy. The highest accuracy is for the class `upperBodyThinStripes, upperBodyVNeck`. The lowest accuracy is 66.0% for the class `upperBodyOther`.

difficult class to accurately measure. On the other hand, the highest accuracy is that of the classes `upperBodyThinStripes` and `upperBodyVNeck`.

For the RAP dataset, similar to the PETA dataset, the obtained results are exceedingly encouraging. The obtained accuracy is 91.1%, while we obtained 92.39%, 91.1%, and 91.56% for the remaining measure precision, recall, and F1-score. The obtained results are a considerable improvement over state of the art. One significant reason for this difference is primarily the large size of the RAP dataset. For the RAP dataset, class-wise accuracy is shown in the Fig. 5. The class `BaldHead` is recognized with a highest accuracy score while the two class that had a low score were that of `Age17-30, Age31-45`. These two classes, naturally, are very difficult to judge, even for experience human observers. Other low performing classes are: `Jacket, OtherAttachments`.

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 wavelet parameters. The method converts the input image into grayscale and then passes it through a series of six mixed-layers blocks that learn the best parameters for the generated Gabor filters. These mixed-layers are a combination of TGW and `3Conv` layers. Output from the last mixed-layer passes through 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 [9]. We eliminate the need to compute pose estimation [24], or construct any hand-crafted features [22]. The discussed results demonstrated superiority over state of the art and justifies the novel use of Gabor wavelet layers.

3. Conclusion

In this paper, we present a novel application of trainable Gabor wavelets to the problem of pedestrian attribute recognition. In contrast to creating offline Gabor filters for image feature extraction, the proposed network learns Gabor wavelets parameters from the data in our deep learning architecture. The network is simple, and has been tested extensively on two of the most challenging publicly available datasets. The results are encouraging and surpass state of the art over many key measures. For the future work, we intend to use these trainable Gabor wavelets with other emerging deep network architectures.

Declarations

Author contribution statement

I.N. Junejo: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

N. Ahmed, M. Lataifeh: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

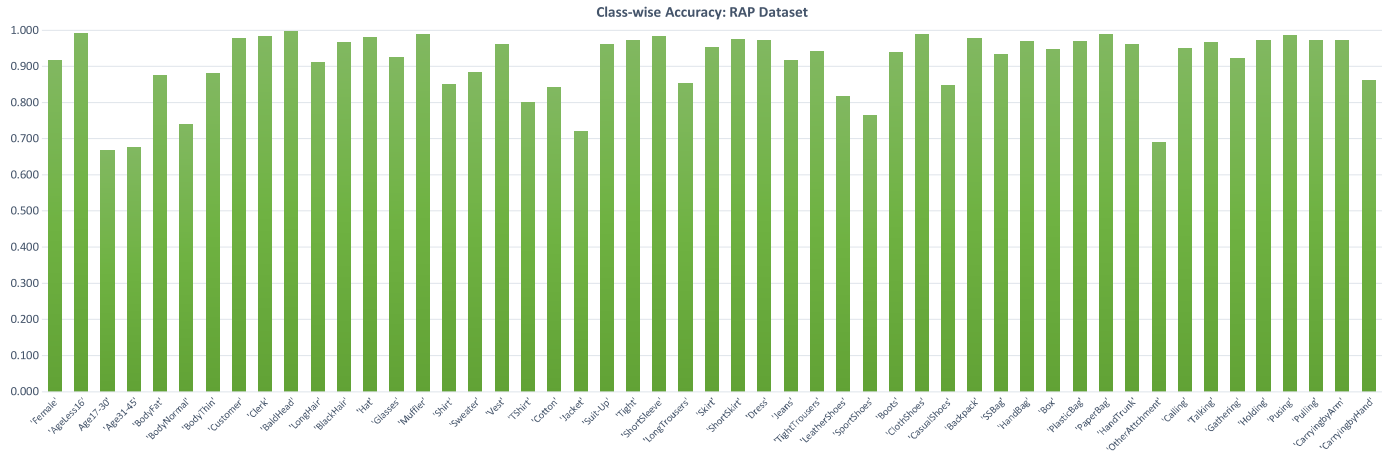

Fig. 5. Class-wise Accuracy - RAP dataset: The lowest accuracy is that of the classes: `Age17-30, Age31-45`. The highest accuracy is for the class `BaldHead`.

Data availability statement

Data included in article/supplementary material/referenced in article.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

[1] L. Cheriet, S. Chenikher, K. Boukari, Fast motion estimation algorithm based on geometric wavelet transform, Int. J. Wavelets Multiresolut. Inf. Process. 17 (2019) 1950018.

[2] F. Raudies, H. Neumann, A bio-inspired, motion-based analysis of crowd behavior attributes relevance to motion transparency, velocity gradients, and motion patterns, PLoS ONE 7 (2013) 1–17.

[3] K. Rahman, N. Abdul Ghani, A. Abdulbasah Kamil, A. Mustafa, M.A. Kabir Chowdhury, Modelling pedestrian travel time and the design of facilities: a queuing approach, PLoS ONE 8 (2013) 1–11.

[4] Yuhao Luo, Dong Yin, An Wang, Wentao Wu, Pedestrian tracking in surveillance video based on modified cnn, Multimed. Tools Appl. 77 (2018) 24041–24058.

[5] A. Nanda, D.S. Chauhan, P.K. Sa, S. Bakshi, Illumination and scale invariant relevant visual features with hypergraph-based learning for multi-shot person re-identification, Multimed. Tools Appl. 78 (2019) 3885–3910.

[6] Y. Yuan, J. Zhang, Q. Wang, Deep Gabor convolution network for person re-identification, Neurocomputing 378 (2020) 387–398.

[7] H. Chandel, S. Vatta, Occlusion detection and handling: a review, Int. J. Comput. Appl. 120 (2015) 0975.

[8] Y. Deng, P. Luo, C.C. Loy, X. Tang, Pedestrian attribute recognition at far distance, in: Proceedings of the 22nd ACM International Conference on Multimedia, MM'14, 2014, pp. 789–792.

[9] D. Li, Z. Zhang, X. Chen, H. Ling, K. Huang, A richly annotated dataset for pedestrian attribute recognition, CoRR, arXiv:1603.07054 [abs], 2016.

[10] D.G. Lowe, Object recognition from local scale-invariant features, in: Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, 1999, pp. 1150–1157.

[11] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, 2005, pp. 886–893.

[12] P. Viola, M. Jones, Robust real-time object detection, Int. J. Comput. Vis. 57 (2001).

[13] H.J. Kwon, H. Koo, J.W. Soh, N.I. Cho, Age estimation using trainable Gabor wavelet layers in a convolutional neural network, in: 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 3626–3630.

[14] X. Wang, S. Zheng, R. Yang, B. Luo, J. Tang, Pedestrian attribute recognition: a survey, arXiv:1901.07474, 2019.

[15] S. Maji, A.C. Berg, J. Malik, Classification using intersection kernel support vector machines is efficient, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[16] Y. Jia, et al., Caffe: convolutional architecture for fast feature embedding, in: Proceedings of the 22nd ACM International Conference on Multimedia, MM'14, 2014.

[17] J. Joo, S. Wang, S. Zhu, Human attribute recognition by rich appearance dictionary, in: 2013 IEEE International Conference on Computer Vision, 2013, pp. 721–728.

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

[19] X. Zhao, et al., Recurrent attention model for pedestrian attribute recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 9275–9282.

[20] J. Zhu, S. Liao, D. Yi, Z. Lei, S.Z. Li, Multi-label cnn based pedestrian attribute learning for soft biometrics, in: 2015 International Conference on Biometrics (ICB), 2015, pp. 535–540.

[21] Y. Zhou, et al., Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization, in: British Machine Vision Conference BMVC 4-7, 2017.

[22] Y. Chen, S. Duffner, A. Stoian, J.-Y. Dufour, A. Baskurt, 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, 2018, pp. 114–122.

[23] S. Liao, Y. Hu, Xiangyu Zhu, S.Z. Li, Person re-identification by local maximal occurrence representation and metric learning, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2197–2206.

[24] D. Li, X. Chen, Z. Zhang, K. Huang, Pose guided deep model for pedestrian attribute recognition in surveillance scenarios, in: 2018 IEEE International Conference on Multimedia and Expo (ICME), 2018, pp. 1–6.

[25] P. Liu, X. Liu, J. Yan, J. Shao, Localization guided learning for pedestrian attribute recognition, in: British Machine Vision Conference 2018, BMVC 2018, 2018.

[26] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: Proceedings of the 32Nd International Conference on International Conference on Machine Learning, vol. 37, ICML'15, 2015, pp. 448–456.

[27] Q. Li, X. Zhao, R. He, K. Huang, Visual-semantic graph reasoning for pedestrian attribute recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 8634–8641.

[28] I.N. Junejo, N. Ahmed, A multi-branch separable convolution neural network for pedestrian attribute recognition, Heliyon 6 (2020).

[29] M. Sarfraz, A. Schumann, Y. Wang, R. Stiefelhagen, Deep view-sensitive pedestrian attribute inference in an end-to-end model, in: British Machine Vision Conference (BMVC), 2017.

[30] M. Saquib Sarfraz, A. Schumann, A. Eberle, R. Stiefelhagen, 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), 2018.

[31] H. An, H. Fan, K. Deng, H.-M. Hu, Part-guided network for pedestrian attribute recognition, in: 2019 IEEE Visual Communications and Image Processing (VCIP), 2019, pp. 1–4.

[32] X. Liu, et al., Hydraplus-net: attentive deep features for pedestrian analysis, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1–9.

[33] N. Sarafianos, X. Xu, I.A. Kakadiaris, Deep imbalanced attribute classification using visual attention aggregation, in: Springer European Conference on Computer Vision, 2018, pp. 708–725.

[34] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using places database., in: Proceedings of the 27th International Conference on Neural Information Processing Systems, vol. 1, NIPS'14, MIT Press, Cambridge, MA, USA, 2014, pp. 487–495.

[35] H. Guo, X. Fan, S. Wang, Human attribute recognition by refining attention heat map, Pattern Recognit. Lett. 94 (2017) 38–45.

[36] W. Li, X. Zhu, S. Gong, Harmonious attention network for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[37] X. Chang, T.M. Hospedales, T. Xiang, Multi-level factorisation net for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[38] J. Wang, X. Zhu, S. Gong, W. Li, Transferable joint attribute-identity deep learning for unsupervised person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[39] J. Si, et al., Dual attention matching network for context-aware feature sequence based person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[40] X. Qian, et al., Pose-normalized image generation for person re-identification, in: The European Conference on Computer Vision (ECCV), 2018.
[41] P. Chikontwe, H.J. Lee, Deep multi-task network for learning person identity and attributes, IEEE Access 6 (2018) 60801–60811.
[42] E. Bekele, W. Lawson, The deeper, the better: analysis of person attributes recognition, in: 14th IEEE International Conference on Automatic Face & Gesture Recognition, FG, 2019.
[43] Qiaozhe Li, Ran He, Xin Zhao, K. Huang, Visual-semantic graph reasoning for pedestrian attribute recognition, in: 33rd AAAI Conference on Artificial Intelligence, AAAI, 2019.
[44] X. Zhao, et al., Recurrent attention model for pedestrian attribute recognition, in: 33rd AAAI Conference on Artificial Intelligence, AAAI, 2019.
[45] A. Alekseev, A. Bobe, Gabornet: Gabor filters with learnable parameters in deep convolutional neural network, in: 2019 International Conference on Engineering and Telecommunication (EnT), 2019, pp. 1–4.
[46] J. Bai, Y. Zeng, Y. Zhao, F. Zhao, Training a v1 like layer using Gabor filters in convolutional neural networks, in: 2019 International Joint Conference on Neural Networks (IJCNN), 2019, pp. 1–8.
[47] F. Ahmed, B.A. Topu, S.M.M. Islam, Hog and Gabor filter based pedestrian detection using convolutional neural networks, in: 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019, pp. 1–6.
[48] S. Luan, et al., Gabor convolutional networks, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, pp. 1254–1262.
[49] F. Chollet, Keras, https://github.com/fchollet/keras, 2015.
[50] G. Guo, G. Mu, Y. Fu, T. Huang, Human age estimation using bio-inspired features, in: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009, 2009, pp. 112–119.
[51] P. Sudowe, H. Spitzer, B. Leibe, Person attribute recognition with a jointly-trained holistic cnn model, in: 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), 2015, pp. 329–337.