IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 26, 2021, accepted February 13, 2021, date of publication February 23, 2021, date of current version March 17, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3061538

Multi-Branch Gabor Wavelet Layers for Pedestrian Attribute Recognition

IMRAN N. JUNEJO
Zayed University, Dubai, United Arab Emirates
e-mail: imran.junejo@zu.ac.ae

ABSTRACT Surveillance cameras are everywhere, keeping an eye on pedestrians as they navigate through a scene. With this context, our paper addresses the problem of pedestrian attribute recognition (PAR). This 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. The problem has rightly attracted attention recently from the computer vision community. In this paper, we adopt trainable Gabor wavelets (TGW) layers and use it with a convolution neural network (CNN). Whereas other researchers are using fixed Gabor filters with the CNN, the proposed layers are learnable and adapt to the dataset for a better recognition. We propose a multi-branch neural network where mixed-layers, a combination of the TGW and convolutional layer, make up the building block of our 3-branch deep neural network. We test our method on publicly available challenging datasets and compare our results with state of the art.

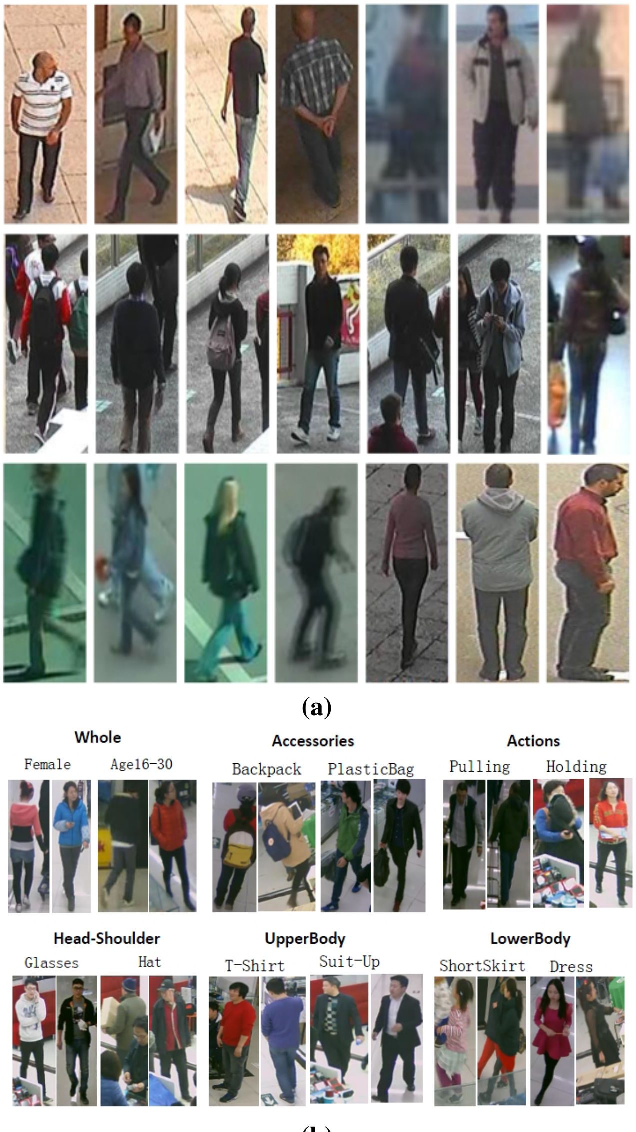
INDEX TERMS Computer vision, pedestrian attribute recognition, deep learning.

I. INTRODUCTION

One of the active areas of research in computer vision is the pedestrian attribute recognition. 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 [1], 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 [2], [3]. 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.

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 the 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 datasets. Figure 1 shows some of the samples from the PEdesTrian Attribute (PETA) [4] and A Richly Annotated Pedestrian (RAP) [5] datasets. 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

The associate editor coordinating the review of this manuscript and approving it for publication was Ikramullah Lali.



(a)

(b)

FIGURE 1. (a) PETA [4] dataset Samples. (b) RAP [5] dataset samples.

blur, many of the attributes are hidden due to severe occlusions. 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 appearance, viewpoints and severe occlusions. After analyzing these datasets, it is observed that visual attributes identification from these images is a difficult task due to the very low quality of the images. Many of the attributes are not completely visible due to 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 [6], HoG [7] or Haar-like features [8] have been employed in the field rigorously. Feature extraction is followed by the attributes classification step. For classification, Support Vector Machines (SVM) [4] 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 few 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, the filter are pre-constructed and then fed as filters on the convolutional network. However, we adopt an approach where the convolutional network is employed to learn the wavelet parameters along with learning the dataset. These Trainable Gabor wavelets (TGW) [9] make up for the backbone of our network. Each TGW accepts a single channel input, with a multi-channel output, and learns the best parameters to generate a 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, refer to as mixed-layer, and make up the building block of our network. The proposed network, shown in Fig. 3, divides an input image into three parts. Each part passes through a separate branches each consisting of 4 mixed-layers. Each branch undergoes a series of fully connected (fc) layers that are connected to the final output layer. The network is simple and is 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 3-branch network that, while learning the Gabor wavelets parameters, and combine the 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.

II. RELATED WORK

In this section we will discuss the works that are related most closely to our method, a detailed survey can be found here [10]. PETA [4] is one of the most widely used pedestrian datasets. While introducing the dataset, the Deng *et al.* [4] used the luminance channel and applied Ensemble of Localized Features (ELF) and Gabor and Schmid filter on it. To address the class imbalance problem they also applied ikSVMs [11] on each attribute separately. They also proposed using the Markov Random Field (MRF) to exploit the context from neighboring images. In their representation, each image is a node and the link between two nodes is determined by the similarity between the images. RAP dataset [5] 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 [12] 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 of ELF. Additionally, they divided the image into multiple blocks (three in their case) to employ a part-based classification scheme. For their work, the parts were comprised of: upper body (torso), lower body, and head and shoulders. Joo *et al.* [13] proposed another approach that also employed part-based recognition. In their work, they first crated Histogram of Oriented Gradient (HoG) features from an image subdivided into multiple overlapping regions. For the attributes classification, they employed a Poselet-based approach [14]. Zhao *et al.* [15] proposed a solution that employed a Recurrent Neural Network (RNN). The authors proposed an end-to-end Recurrent Convolutional (RC) and Recurrent Attention (RA) models. RC model mines the correlations among different attribute groups, while the intra-group attention correlation and intra group spatial locality is used by the RA model to improve the performance and robustness of pedestrian attribute recognition. However, their network has a deep architecture, hence the number of parameters is quite large. In another part-based approach, Zhu *et al.* [16] 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.* [17] first extracted mid-level features from detection layers using GoogLeNet. They localized the pedestrian attributes by fusing and clustering the activation maps of the detection layers. Only the image labels are used to train the detected layers in order to learn the relationship between the mid-level features and the pedestrian attributes. For training a max-pooling based weakly-supervised object detection technique is employed. Chen *et al.* [18] proposed a part-based network that combined LOMO features [19] 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. Li *et al.* [20] used pedestrian body structure knowledge and proposed a pose-guided model. In the first step, the model computes the transformation parameters to estimate the pose from the image. Based on the pose information it then localizes the body parts. Final attribute recognition is estimated by fusing multiple features. Another parts localization method is proposed by Liu *et al.* [21]. They proposed a Localization Guide Network (LGNet) that uses a CNN model based on Inception-v2 [22] for feature extraction. Afterward, a global average pooling layer (GAP) is adopted to extract global features. The fusion of global and local features is used to obtain the pedestrian attributes classification. Li *et al.* [23] presented a visual semantic graph based approach that used ResNet-50 to for the pedestrian images feature extraction. Junejo and Ahmed [24] also presented a multi-branch approach using different color space input. The proposed network contains a large number of parameters because it had more than fifty layers.

Sarfraz *et al.* [25] 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 [26] 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 [27] 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.* [28]. HydraPlus-Net is an Inception-based network and aggregates feature layers from multi-directional attention modules for the final feature representation. Sarafianos *et al.* [29] 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 different scales to obtain a better feature representation. Another end-to-end method for person attribute recognition that uses Class Activation Map (CAM) network [30] to refine attention heat map is proposed by Guo *et al.* [31]. 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 [32]. 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 [33]. 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 attribute-semantic feature representation is proposed by [34]. Si *et al.* [35] 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 [36]. They learn pose invariant deep person re-identification features using synthesized images. A deep CNN based method to learn partial descriptive features for efficient person feature representation is presented in [37]. They employed a pyramid spatial pooling module and reported an improvement of 2.71% on the PETA dataset over [25]. Reference [38] improved over [25] by employing a deeper network based on a context sensitive framework. The proposed network improved generalization and classification accuracy by creating a richer feature sets using deeper residual networks (ResNet) and achieved the best in class results on attribute recognition datasets. Reference [23] presented a visual semantic graph reasoning framework that modeled spatial and attribute relationships using two types of graphs. For reasoning, they employed Graph Convolutional Network that encapsulates the spatial relationship between local regions of the image and the potential semantic relationship of the attributes. Reference [15] used Recurrent Attention (RA) and Recurrent Convolutional (RC) to present a dual model approach for pedestrian recognition. The RC model employed a Convolutional-LSTM model to establish the correlations between the different groups of attributes. To improve the overall robustness, the RA model used both local attention correlation and global spatial locality.

Using Gabor wavelets with CNNs have received a tremendous attention as well [9], [39]–[41]. Reference [39] 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. Reference [40] 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 [41] 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, [9] 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 two challenging datasets and show a considerable improvement over state of the art.

III. 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.

A. GABOR WAVELET LAYER

We make use of the Trainable Gabor wavelets (TGW) layer as proposed by Kwon *et al.* [9] (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.

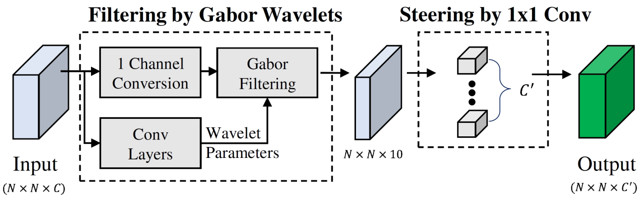


FIGURE 2. Trainable Gabor Wavelet (TGW) layer [9]: Inputs and outputs are multichannel. A neural network is used to generate Gabor wavelet hyperparameters. These generated Gabor filters are then applied to the input. 1 × 1 convolution layer is added to enable the steerability of the Gabor wavelets.

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, 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 [9]. 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,

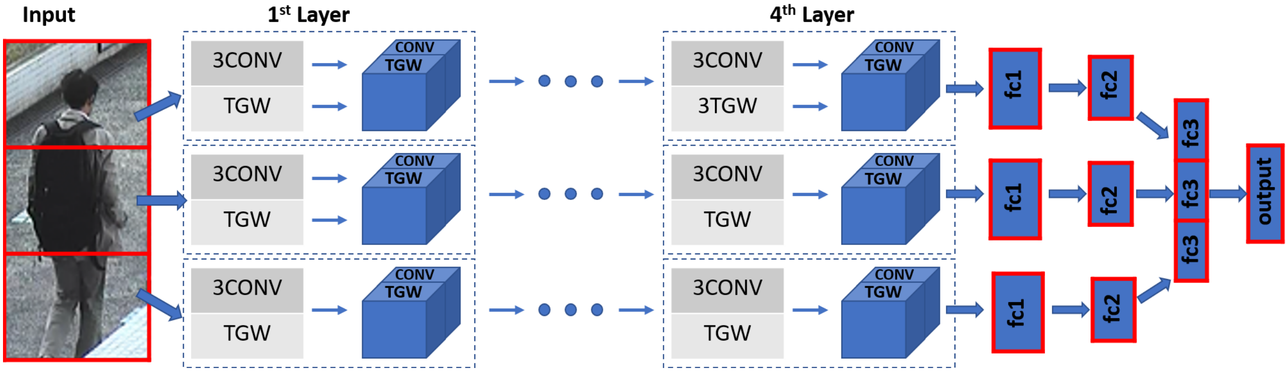


FIGURE 3. Our Approach: The proposed method divides the input image into three parts. For each branch, the network contains 4 layers that are a mix between TGW and 3Conv layer (mixed-layers). The output of each branch is followed by three fc layers. Size of the last layer of the network matches the number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

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° [9].

B. 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 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 3×3 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.

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	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

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.

IV. 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 3Conv 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 ($F1$) [5].

A. 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.

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 [4]				**RAP [5]**			
	Acc	*Prec*	*Rec*	*F1*	*Acc*	*Prec*	*Rec*	*F1*
Chen et. al. [20]	75.07	83.68	83.14	83.41	62.02	74.92	76.21	75.56
Li et. al. [5]	–	–	–	–	63.67	76.53	77.47	77.00
Sudowe et. al. [46]	73.66	84.06	81.26	82.64	62.61	80.12	72.26	75.98
Liu et. al. [17]	74.62	82.66	85.16	83.40	53.30	60.82	78.80	68.65
Sarfaraz et. al. [25]	77.73	86.18	84.81	85.49	67.35	79.51	79.67	79.59
Li et. al. [28]	76.13	84.92	83.24	84.07	65.39	77.33	78.79	78.05
ours	**79.35**	**86.24**	79.45	81.48	**90.93**	92.59	90.9	91.5

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 [20], [25] 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 fare 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 [25]. We adopted the weighted-cross entropy loss function [20] in order to mitigate the class imbalance problem. Similarly, following other researchers, images are resized to an image resolution of 144×48.

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.

B. SETUP

For deep learning, we adopted the KERAS [42] 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.

C. 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 [43]: 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.

D. 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 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`.

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: `Cotton, 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 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 [5]. We also do not have to compute pose estimation [20], or construct any hand-crafted

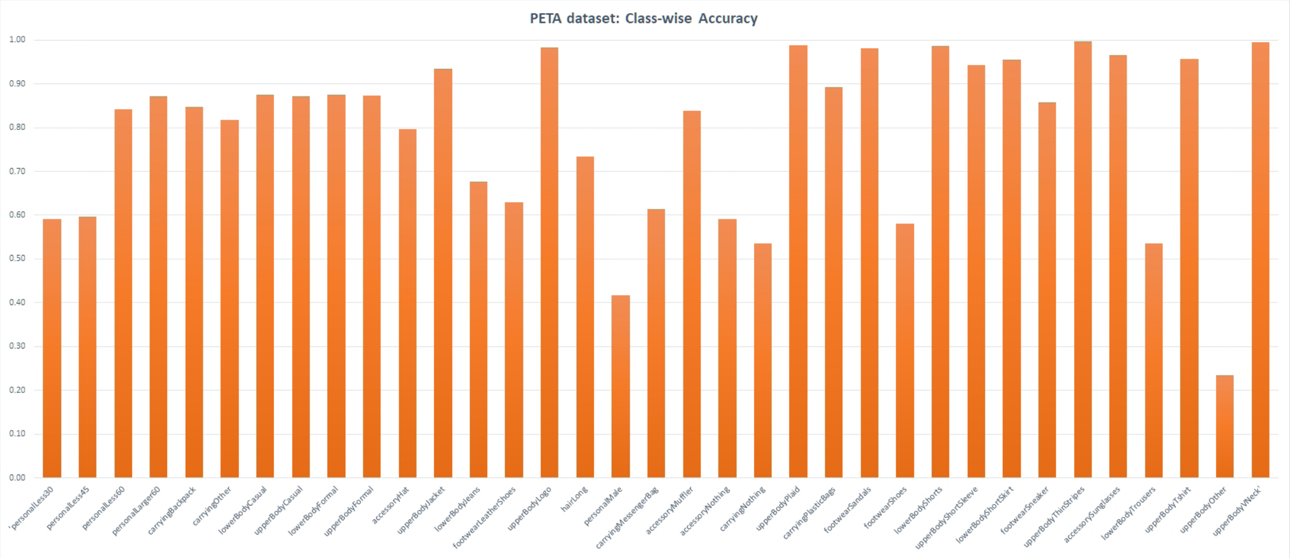


FIGURE 4. Class-wise Accuracy - PETA dataset: the figure shows the obtained class-wise accuracy. The highest accuracy is for the class `upperBodyThinStripes,upperBodyVNeck`**. The lowest accuracy is 23.4% for the class** `upperBodyOther`**.**

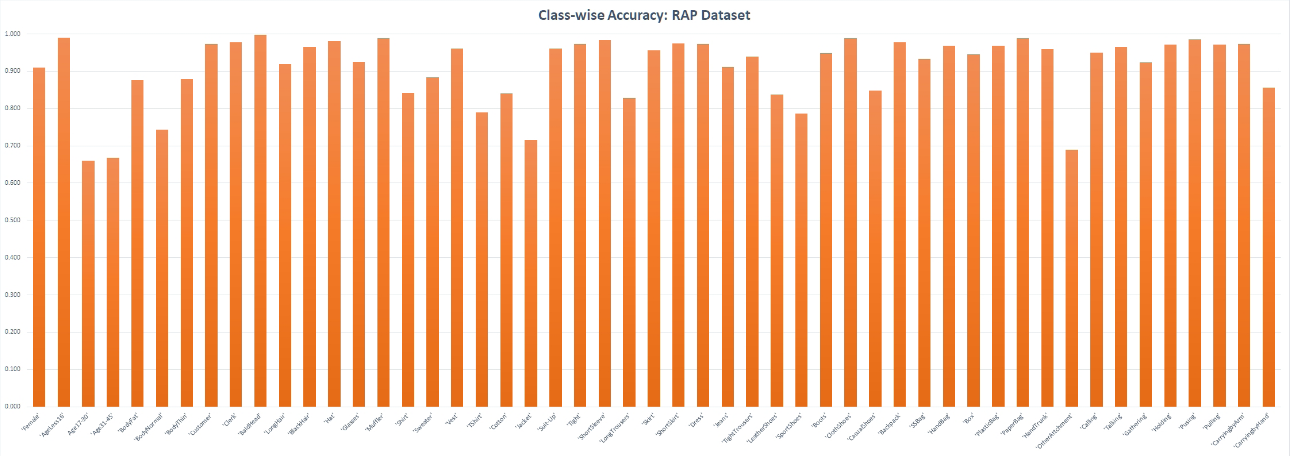


FIGURE 5. Class-wise Accuracy - RAP dataset: The lowest accuracy is that of the classes: `Age17-30, Age31-45`**. The highest accuracy is for the class** `BaldHead`**.**

features [18]. Our results are an improvement over state of the art and clearly justifies the use of Gabor wavelet layers.

V. CONCLUSION

We propose a novel multi-branch neural network. Our key contribution is using trainable Gabor wavelets (TGW) for the pedestrian attribute recognition problem. The input image is divided into three parts and each part is processed through three branches of the network. Each branch contains mixed-layers that are capable of learning the Gabor wavelet parameters. This is very crucial, as filters are learned from the structure of the dataset itself. We demonstrate the workings of our network on two of the most challenging public datasets and show very encouraging results. For future work, we intend to further investigate Gabor wavelets for the PAR problem with different network architectures.

REFERENCES

[1] F. Raudies and H. Neumann, "A bio-inspired, motion-based analysis of crowd behavior attributes relevance to motion transparency, velocity gradients, and motion patterns," *PLoS ONE*, vol. 7, no. 12, Dec. 2013, Art. no. e53456.

[2] K. Rahman, N. A. Ghani, A. A. Kamil, A. Mustafa, and M. A. K. Chowdhury, "Modelling pedestrian travel time and the design of facilities: A queuing approach," *PLoS ONE*, vol. 8, no. 5, May 2013, Art. no. e63503.

[3] A. Nanda, D. S. Chauhan, P. K. Sa, and S. Bakshi, "Illumination and scale invariant relevant visual features with hypergraph-based learning for multi-shot person re-identification," *Multimedia Tools Appl.*, vol. 78, no. 4, pp. 3885–3910, Feb. 2019.

[4] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 789–792.

[5] D. Li, Z. Zhang, X. Chen, H. Ling, and K. Huang, "A richly annotated dataset for pedestrian attribute recognition," *CoRR*, vol. abs/1603.07054, Mar. 2016.

[6] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Dec. 1999, pp. 1150–1157.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 886–893.

[8] P. Viola and M. Jones, "Robust real-time object detection," *Int. J. Comput. Vis.*, vol. 57, pp. 34–47, Jul. 2001.

[9] H. J. Kwon, H. Il Koo, J. W. Soh, and N. Ik Cho, "Age estimation using trainable Gabor wavelet layers in a convolutional neural network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 3626–3630.

[10] X. Wang, S. Zheng, R. Yang, B. Luo, and J. Tang, "Pedestrian attribute recognition: A survey," 2019, *arXiv:1901.07474*. [Online]. Available: http://arxiv.org/abs/1901.07474

[11] S. Maji, A. C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[12] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 675–678.

[13] J. Joo, S. Wang, and S.-C. Zhu, "Human attribute recognition by rich appearance dictionary," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 721–728.

[14] L. Bourdev, S. Maji, and J. Malik, "Describing people: A poselet-based approach to attribute classification," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1543–1550.

[15] X. Zhao, L. Sang, G. Ding, J. Han, N. Di, and C. Yan, "Recurrent attention model for pedestrian attribute recognition," in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2019, vol. 33, no. 1, pp. 9275–9282.

[16] J. Zhu, S. Liao, D. Yi, Z. Lei, and S. Z. Li, "Multi-label CNN based pedestrian attribute learning for soft biometrics," in *Proc. Int. Conf. Biometrics (ICB)*, May 2015, pp. 535–540.

[17] Y. Zhou, K. Yu, B. Leng, Z. Zhang, D. Li, and K. Huang, "Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization," in *Proc. Brit. Mach. Vis. Conf.*, 2017.

[18] Y. Chen, S. Duffner, A. Stoian, J.-Y. Dufour, and A. Baskurt, "Pedestrian attribute recognition with part-based CNN and combined feature representations," in *Proc. 13th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2018, pp. 114–122.

[19] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2197–2206.

[20] D. Li, X. Chen, Z. Zhang, and K. Huang, "Pose guided deep model for pedestrian attribute recognition in surveillance scenarios," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.

[21] P. Liu, X. Liu, J. Yan, and J. Shao, "Localization guided learning for pedestrian attribute recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2018.

[22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, vol. 37, Jul. 2015, pp. 448–456.

[23] Q. Li, X. Zhao, R. He, and K. Huang, "Visual-semantic graph reasoning for pedestrian attribute recognition," in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2019, vol. 33, no. 1, pp. 8634–8641.

[24] I. N. Junejo and N. Ahmed, "A multi-branch separable convolution neural network for pedestrian attribute recognition," *Heliyon*, vol. 6, no. 3, Mar. 2020, Art. no. e03563.

[25] M. S. Saquib, A. Schumann, Y. Wang, and R. Stiefelhagen, "Deep view-sensitive pedestrian attribute inference in an end-to-end model," in *Proc. Brit. Mach. Vis. Conf.*, 2017.

[26] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 420–429.

[27] H. An, H. Fan, K. Deng, and H.-M. Hu, "Part-guided network for pedestrian attribute recognition," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2019, pp. 1–4.

[28] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang, "HydraPlus-net: Attentive deep features for pedestrian analysis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1–9.

[29] N. Sarafianos, X. Xu, and I. A. Kakadiaris, "Deep imbalanced attribute classification using visual attention aggregation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 708–725.

[30] R. Raturi, "Adapting deep features for scene recognition utilizing places database," in *Proc. 2nd Int. Conf. Inventive Commun. Comput. Technol. (ICICCT)*, Apr. 2018, pp. 487–495.

[31] H. Guo, X. Fan, and S. Wang, "Human attribute recognition by refining attention heat map," *Pattern Recognit. Lett.*, vol. 94, pp. 38–45, Jul. 2017.

[32] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2285–2294.

[33] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2109–2118.

[34] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2275–2284.

[35] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang, "Dual attention matching network for context-aware feature sequence based person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5363–5372.

[36] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, and X. Xue, "Pose-normalized image generation for person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 650–667.

[37] P. Chikontwe and H. J. Lee, "Deep multi-task network for learning person identity and attributes," *IEEE Access*, vol. 6, pp. 60801–60811, 2018.

[38] E. Bekele and W. Lawson, "The deeper, the better: Analysis of person attributes recognition," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–8.

[39] A. Alekseev and A. Bobe, "GaborNet: Gabor filters with learnable parameters in deep convolutional neural network," in *Proc. Int. Conf. Eng. Telecommun.*, Nov. 2019, pp. 1–4.

[40] J. Bai, Y. Zeng, Y. Zhao, and F. Zhao, "Training a v1 like layer using Gabor filters in convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.

[41] S. Luan, B. Zhang, S. Zhou, C. Chen, J. Han, W. Yang, and J. Liu, "Gabor convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2018, pp. 1254–1262.

[42] F. Chollet. (2015). *Keras*. [Online]. Available: https://github.com/fchollet/keras

[43] G. Guo, G. Mu, Y. Fu, and T. S. Huang, "Human age estimation using bio-inspired features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 112–119.

[44] P. Sudowe, H. Spitzer, and B. Leibe, "Person attribute recognition with a jointly-trained holistic CNN model," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 329–337.

IMRAN N. JUNEJO was a Researcher with the National Institute for Research in Computer Science and Automation, a unit of the French National Institute for Research in Computer Science and Control, from 2007 to 2009. From 2009 to 2017, he was an Associate Professor with the University of Sharjah. From August 2017 to August 2020, he worked as an Associate Professor and a Full Professor with the Institute of Business Administration, Karachi. He is currently an Associate Professor with Zayed University, Dubai, United Arab Emirates. He is also working with Laptev and Patrick Perez at the Vista Group. His primary areas of research are computer vision and machine learning. His current focus of research is application of deep learning to the classical computer vision problems. His research interests include human action recognition from arbitrary views, camera calibration, crowd modeling and analysis, path modeling, video surveillance, scene understanding, and event detection.