3-1-2020

A multi-branch separable convolution neural network for pedestrian attribute recognition

Imran N. Junejo
Zayed University

Naveed Ahmed
University of Sharjah

Heliyon 6 (2020) e03563

Contents lists available at ScienceDirect

Heliyon

www.elsevier.com/locate/heliyon

Research article

A multi-branch separable convolution neural network for pedestrian attribute recognition

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

[a] Zayed University, Dubai, United Arab Emirates
[b] University of Sharjah, United Arab Emirates
[c] Institute of Business of Administration, Karachi, Pakistan

ARTICLE INFO

Keywords:
Computer science
Computer Vision
Image processing
Deep learning
Pedestrian attribute recognition

ABSTRACT

Video surveillance applications have made great strides in making the world a safer place. Extracting visual attributes from a scene, such as the type of shoes, the type of clothing, carrying any object or not, or wearing any accessory etc., is a challenging problem and an efficient solution holds the key to a great number of applications. In this paper, we present a multi-branch convolutional neural network that uses **d**epthwise **s**eparable **c**onvolution (DSC) layers to solve the pedestrian attribute recognition problem. Researchers have proposed various solutions over the years making use of convolutional neural networks (CNN), however, we introduce DSC layers to the CNN for the problem of pedestrian attribute recognition. In addition, we make a novel use of the different color spaces and create a 3-branch CNN, denoted as 3bCNN, that is efficient, especially with smaller datasets. We experiment on two benchmark datasets and show results with improvement over the state of the art.

1. Introduction

One relatively recent problem that has peeked research interest lately is that of pedestrian attribute recognition. The goal is to identify visual attributes from images, such as age group, gender, clothing style, footwear etc. The problem has a number of applications, mainly in the areas of person identification or visual surveillance, or in the domains for business intelligence by means of video analytics etc. Using the recognized visual attributes, one can search for the suspects in a criminal database. It is a challenging problem because of a number of factors. As there are a large number of variations in these visual attributes, it is one of the main factors that makes this problem difficult to solve. As an example, due to varying lighting conditions, same type of clothing can appear completely different or vice versa. Additionally, weather conditions play an important role in how an attribute will appear to the camera, for example, depending on the conditions, it will be very difficult to distinguish between dark blue and black colors. Furthermore, a number of visual attributes can be completely or partially occluded due to the camera orientation, such as a backpack that can be hidden from the view, or a scarf or the hat that may not appear in the image due to the viewing direction or self occlusions. These issues highlight the fact that a very high intra-class variation exists for the same attributes depending on the number of conditions that exist at the time of image or video acquisition. As this work focuses on the image and video data, the distance of the pedestrian from the camera poses another challenge. As a common practice, the surveillance cameras are typically installed at quite a distance, therefore the acquired image is not able to capture the pedestrian with a good detail. Depending on the image resolution and the distance, the size of the object can be very small, and the poor image quality results in very few pixels assigned to each attribute, e.g. shoes, hats or a backpack can only be a small number of pixels in an already low resolution image. Due to self occlusions, body parts are not always fully visible, and lack in important visual data because of a very low image quality. Some of the sample images from the PEdesTrian Attribute (PETA) and A Richly Annotated Pedestrian (RAP) datasets are shown in Fig. 1. It can be seen that the images are of a very low quality due to a number of reasons: attributes are not visible to due to severe occlusions, the image quality is quite low, some of the images show a significant blur due to acquisition problems.

Most of the current works that try to solve the visual attribute recognition problem propose a two step solution comprising of feature extraction followed by the attribute classification. Derived features, such as SIFT [1], HoG [2] or Haar-like feature [3], have been predominantly employed for the feature representation in the earlier works. These derived features, which are employed in most of the computer vision solutions over the past two decades, not only need a very high

* Corresponding author at: Zayed University, Dubai, United Arab Emirates.
E-mail address: ijunejo@cs.ucf.edu (I.N. Junejo).

https://doi.org/10.1016/j.heliyon.2020.e03563
Received 6 August 2019; Received in revised form 12 December 2019; Accepted 5 March 2020

Fig. 1. (a) Some of the samples from the PEdesTrian Attribute (PETA) dataset. It is one of the largest dataset that covers more than 60 attributes in 19,000 images of different resolution. A total of 8,705 persons are included in this dataset that are captured from real-world surveillance camera systems. The dataset is very challenging, due to scene settings and the acquisition setup. It can be seen that the images are of a very low quality due to a number of reasons: attributes are not visible to due to severe occlusions, the image quality is quite low, some of the images show a significant blur due to acquisition problems. (b) Some of the samples from the A Richly Annotated Pedestrian (RAP) dataset. This dataset is acquired from more than one viewpoints and covers around 72 attributes over 41 thousand images. The dataset has a large number of variation in viewpoints, pedestrian appearance, and severe occlusions.

domain knowledge but also require fine tuning for an accurate representation. After the feature representation step, most of the solutions employed Support Vector Machines (SVM) for the features classification [4].

SVM are increasingly replaced by the convolutional neural networks (CNNs) in the last five years. CNNs have been demonstrated to outperform many of the previous approaches for both tasks of attribute learning or image classification. Specifically, we also employ a CNN, but instead of using the regular 2D convolution layers, we adopt the depthwise separable convolution (DSC) layer as proposed in the Xception [5] framework. DSC layers have been used in various applications due to its efficient learning capabilities and reduced parameter set [6, 7, 8]. As shown Fig. 2, each input channel (3 in our case) is treated independently from other channels. The split channels are convolved with a 3×3 spatial filter. The output channels are concatenated and then convolved with a $1 \times 1 \times n$ filter, where n matches the depth of our channels. This process reduces the number of parameters for our network considerably. Fewer computations are performed, and results in speeder and smaller models. Specially, when the data available is not sufficiently large, DSC layers have been shown to learn a better data representation and learns better performing models [5].

Based on our previous work [9], the main contributions of the proposed method are:

- To the best of our knowledge, this is the first work to introduce depthwise separable convolution neural networks for the problem of pedestrian attribute recognition.
- The proposed multi-layered network is trained efficiently with a smaller number of parameters.
- We make novel use of the color spaces for training our network.
- The proposed method is demonstrated to have better recognition results than the state of the art on two of the most challenging public datasets.

2. Related work

Some of the works mostly closely related to our method are discussed below. PETA dataset was introduced by [4]. They make use of

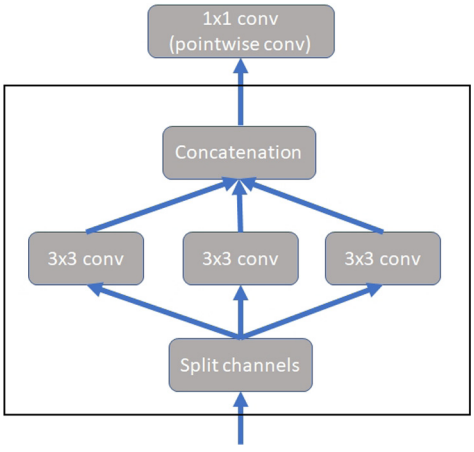

Fig. 2. Depthwise Separable Convolution. (Source: [5]).

Gabor and Schmid filter on the luminance channel, in addition to the Ensemble of Localized Features (ELF). ikSVMs are applied on each attribute separately to avoid the class imbalance problem. In order to exploit the context from neighboring images, they propose using the Markov Random Field (MRF), where an image is represented by a node and the similarity between the images determines the link between any two nodes. Based on Caffe [10] framework employing two CNN models, [11] analyze the influence of viewpoint variations, occlusions as well as body parts on the overall classification by adopting the ELF, in addition to training the SVMs. In addition, a part-based classification is performed where an image is divided into three blocks: the upper body, head and shoulders, and the lower body parts. Another part-based recognition approach was proposed by [12], in which an image first is subdivided into overlapping regions that are used as an input to create a Histogram of Oriented Gradient (HoG) features. They then use a Poselet-based approach [13] for the attributes' classification. Dividing the human body part into 15 parts, [14] train a separate CNN for each part where the weight of a CNN is based on the contribution of a particular attribute. [15] makes use of the GoogLeNet to extract mid-

level features from detection layers. Activation maps of these detected layers are fused, and clustering is performed to localize pedestrian attributes. In order to learn the relationship between the mid-level features and the attributes, these detected layers are trained using image labels only by adopting a max-pooling based weakly-supervised object detection technique. Combining CNN extracted features with LOMO features, a part-based network is proposed by [16]. LOMO features, based on HSV histograms and Scale-Invariant Local Ternary Patterns, are texture and color descriptors shown to be illumination-invariant. [17] propose a pose-guided model that uses pedestrian body structure knowledge. First, the model estimates the pose from the image by computing the transformation parameters, and then they perform body part localization. Multiple features are fused to estimate the final attribute recognition. A Localization Guide Network (LGNet) is proposed by [18], able to localize areas corresponding to different attributes. They use an Inception-v2 [19] as their base CNN model for feature extraction and adopt a global average pooling layer (GAP) to extract global features. The pedestrian attributes classification is obtained by the fusion of global and local features. Some recent works focus on low-dimensional feature mappings, or imposing regularization constraint to learn rotation-invariant features as well [20, 21, 22, 23].

CNN based network (VeSPA) is proposed by [24]. This end-to-end network estimates four pose categories. A separate part of the network learns each of these pose-specific attributes. They aim to show that pedestrian attribute recognition is greatly influenced by coarse body pose information. Later, they extended their work and proposed a modified method [25] by adding a ternary view classifier. In contrast to their earlier work they propose a global weighting solution for feature maps before the final embedding. In addition to performing attribute recognition, HydraPlus-Net [26] also performs person re-identification. This Inception-based network employs multi-directional attention modules. Features layers from these modules are aggregated for the final feature representation. A multi-branch network is proposed by [27]. In order to address the class imbalance problem, a simple weight scheme is adopted. For guiding the network to crucial body parts, visual attention masks are extracted at different stages of the network and fused at different scales. This results in a better feature representation. Another end-to-end method to use and refine attention map for person attribute recognition is proposed by [28]. The method uses Class Activation Map (CAM) network [29] and improves the recognition by refining the attention heat map, which is obtained by using CAM and identifies the areas of different image attributes. A joint learning approach for person re-identification using Harmonious Attention CNN (HA-CNN) is proposed by [30]. They perform joint-learning of soft pixels attention and hard regional attention using HA-CNN for the simultaneous optimization of feature representations. A Multi-Level Factorization Net (MLFN) for person re-identification is proposed by [31], in which the visual appearance of a person is factored into latent discriminative factors at multiple semantic levels without manual annotation. A Transferable Joint Attribute-Identity Deep Learning (TJ-AIDL) model for person re-identification is proposed by [32]. This model allows for simultaneously learning of an attribute-semantic and identity discriminative feature representation. Another joint learning end-to-end network for person re-identification, Dual ATtention Matching network (DuATM), is proposed by [33]. This method performs a simultaneous attentive sequence comparison and context-aware feature sequences learning.

A pose-normalized person re-identification framework based on Generative Adversarial Network is proposed by [34]. They use synthesized images to learn deep re-identification features free from influence of pose variation. A pyramid spatial pooling module for efficient person feature representation is employed to learn partial discriminative features using a deep CNN in [35]. They report an improvement of 2.71% on the PETA dataset over [24]. A context sensitive framework using a deeper network to improve classification accuracy and generalization is presented by [36]. They improve over [24] by creating a richer feature sets using deeper residual networks (ResNet) that could achieve the best in class results on attribute recognition datasets. A visual semantic graph reasoning framework is proposed by [37]. The framework contains two types of graphs for modeling spatial relationships and attribute relationships. Graph Convolutional Network is used for reasoning, which can describe the spatial relationship between local regions of the image and the potential semantic relationship of the attributes. A dual model approach using Recurrent Convolutional (RC) and Recurrent Attention (RA) for pedestrian recognition is proposed by [38]. A Convolutional-LSTM model is employed by the RC model to find the correlations between different attribute groups. The RA model makes use of the global spatial locality and local attention correlation to improve the overall robustness. Correlation between attributes is explored by [39]. Their multi-branch network collects context information to compute attribute probabilities. This information is then fused with the results from each branch for improved results.

In contrast to the above approaches, we propose a simple convolutional neural network, that contains DSC layers `dscl_fc_based`, and its variation `dscl_conv_based`. The proposed 3-branch convolutional neural network (3bCNN) makes use of different color space. The input to each branch is an image in YCrCb, L*a*b, HSV color spaces, respectively. The output from each branch is concatenated and flows through a series of fully connected layers before the network output. Our approach reduces the parameter set, learns the model very well and efficiently, especially on small datasets, and produces comparable, if not better, results.

3. Main approach

We describe the proposed method next. We start with how we represent our features. This is then followed by a standard pre-processing step. We describe the proposed deep learning framework thereafter.

3.1. Pre-processing

In this work, we propose 3bCNN to solve the problem of pedestrian attribute recognition. RGB has been the color space of choice for computer vision researchers over the years. Other well-known colors spaces include YCrCb, HSV an L*a*b. While separating the chroma and the luminance component, each color space has its set of advantages. For the task of color compression for video, YCrCb has shown to be very effective. Being device independent, especially for color detection, HSV has shown to be very useful (e.g. skin or hair detection). We are aiming for this strong feature of HSV to assist us in our problem of attribute recognition. L*a*b has shown to be very close to the human color perception, thus colors in L*a*b can be compared using Euclidean distance and the results have been shown to match the human perception of the color as well. Using these color spaces (i.e. HSV, YCrCb and L*a*b), we train our network and present the results below. The image is resized to a resolution of 144×48.

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 centering the data around the origin. Next step involves **normalization**: We compute the standard deviation separately for each color space and the image data is divided by this value. As discussed above, DSC layers are suitable especially for the cases where data size is limited. Hence, we do not resort to data augmentation to train our network.

3.2. Attribute learning

Many of the existing approaches treat each attribute independent of the other attributes. However, this might not always be the case. A person wearing formal dress is more likely to wear formal shoes etc. Therefore, instead of learning each attribute independently, we jointly

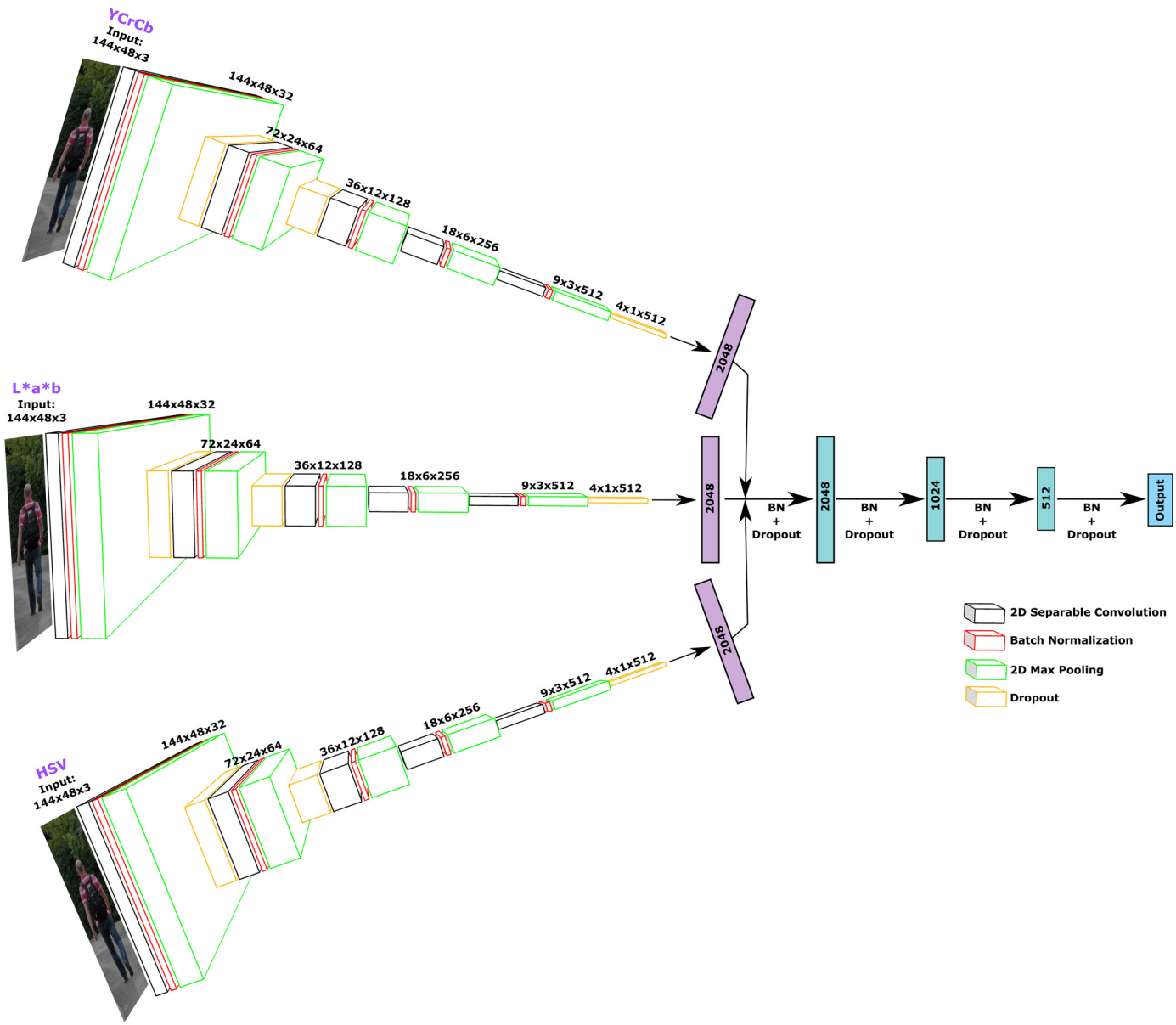

Fig. 3. Proposed Network: We train the network on YCrCb, L*a*b and HSV images, as mentioned above. Input is an image of size $144 \times 48 \times 3$. We use depth-wise separable convolution layers, followed by batch-normalization, max-pooling and dropout layers in succession. The size of output layer matches the number of classes in the dataset.

model the attributes and inter-attribute relationships using deep learning. The proposed network is shown in Fig. 3. We intentionally design the network to maximize the benefit of using multiple color spaces. The proposed network is a three branch network. The input to each branch is an image of resolution 144×48. Each branch gets an input image with a different color space: YCrCb, L*a*b and HSV. The output layer's size matches the number of attributes we aim to recognize [16].

The network architecture is shown in Fig. 3. An input image is first applied with a single stride 2D DSC layer through each branch of the network. Appropriate padding is applied to keep the output size same as the input image size. The output of this layer has depth of 32 channels. We use `LeakyReLU` as the activation function throughout the network, except at the final output. The network then applies batch-normalization (BN), to minimize over-fitting and avoid any covariate shift, followed by a max-pooling layer with pool size [2 2] and a stride [2 2]. The output of this layer is a reduced sized activation map (72×24). Our network follows the pattern: `depthwise_conv -> LeakyReLu -> BN -> max-pool` five times, thus each of the three branches is composed of five blocks. The depth of the network in the first block is 32 channels, 64 in the second, 128 in the third, 256 and 512 in the fourth and the final block, respectively. Thus, depth of the convolution layers increases by a factor of two, with the final size of each branch being $4 \times 1 \times 512$. This is flattened to a size of 2048. Thus, the output of each YCrCb, L*a*b and the HSV layer is of size 2048. These three layers are then concatenated to a layer of size 6144. The network then includes three fully connected (fc) layers of size 2048, 1024 and 215, respectively, followed by the final output layers. Each fc layer sandwiches a BN and a dropout layer with `LeakyReLu` as the activation function. The size of the final layers is equal to the number of pedestrian attributes for a particular dataset. (in our case 35 for PETA and 51 for RAP dataset). We denote this network architecture as **`dscl_fc_based`**.

As a variation to the above described network, we propose **`dscl_conv_based`**. In this architecture, the output of the YCrCb, L*a*b and the HSV layer is concatenated to construct a 3d tensor ($4 \times 512 \times 3$). Treating this as a color image of size 4×512, we apply

Table 1
A comparison of training time and the number of parameters for the proposed **dscl_conv_based**, **dscl_fc_based** and **conv2d_based** (the same network using regular 2D convolutional layers). For this comparison, training is performed on 100 epochs.

	PETA [4]/RAP [11]	
	Time (sec)	# params
dscl_conv_based	1302	549,366
dscl_conv_based using conv2d	1806	4,717,268
dscl_fc_based	2605	15,789,844
dscl_fc_based using conv2d	2774	19,937,571

the sequence: `depthwise conv -> LeakyReLu -> BN -> max-pool` three times. The output is flattened to the size of 128 neurons, followed by the dropout layer (prob = 0.35). The size of the final output layer equals the number of parameters being tested. This architecture dramatically reduces the number of parameters, while still maintaining good accuracy results.

The proposed network is an end-to-end system using depthwise separable convolutional layers. The total number of trainable parameters for our network are $\sim 549,366$, compared to the ~ 19.9 million parameters for the same network using regular convolutional layers. We chose the particular DSC layers for our 3bCNN for its demonstrated versatility, and efficient learning, reduced number of parameters, and an excellent performance on data of small size.

4. Evaluation

As described above, the input to the **dscl_fc_based** network is a YCrCb, L*a*b and an HSV image. The input passes first through the DSC layers. The depth of this initial set of layers is 32. The size of the output layer matches the number of classes in our dataset. The network (cf. Fig. 3) contains a series of `BN` and `max-pool` layers. Before the output layer, we have a three fc layers (`fc1 -> BN -> fc2 -> BN -> fc3 -> BN -> dropout -> output`). `LeakyReLU` is used as the activation function for the fc layers. The final output layer applies the `sigmoid` activation function. Similarly, for the **dscl_fc_based** network, we apply the concatenated three branches with: `depthwise_conv -> BN -> max-pool -> depthwise_conv -> BN -> max-pool -> depthwise_conv -> dropout -> output`. Here also `LeakyReLU` is used as the activation function. Table 1 describes the difference in the number of parameters for each network, and the time needed for training for each network as well.

4.1. Dataset

In order to demonstrate the working of our the proposed method, we experiment on the popular and standard publicly available datasets: PETA [4] and the RAP dataset [11]. PETA contains 19,000 images collected from real-time surveillance cameras. Originally, the dataset contains 61 binary attributes and 4 multi-class attributes. The images contained are of resolution ranging from 17×39 to 169×365. Images have been collected from 10 other publicly available datasets. The RAP dataset contains 41,585 pedestrian samples collected from multi-camera surveillance systems from around 26 different camera setups. Image resolution ranges from 36×92 to 344×554. Each image is annotated with 72 attributes, occlusions, viewpoints, and body parts. Each image in the dataset is annotated independently. The dataset contains images annotated with four types of viewpoints, based on full body direction.

Following previous researchers [17, 24], we report results on 35 attributes for the PETA dataset and for the RAP dataset [24], similarly, we report results on 51 attributes. Inline with the previous research, we divide the data in to 5 random splits: For training, we allocate 9,500 samples. For validation, 1,900 samples are allocated and for testing, 7,600 samples are allocated for the PETA dataset. The ratio for training, and testing the RAP dataset is: $33,268 : 8,317$ images. In order to address the imbalance in the data, we adopt a weighted-cross entropy loss function [17].

4.2. Setup

For deep learning, we adopted the KERAS [40] 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.

4.3. Implementation details

We train for 250 epochs while using the `LeakyReLU` activation function with $\alpha = 0.0001$ for all layers. Using the `Adam` update optimizer, the following parameters were set: learning rate = 0.001, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We added the dropout layers to prevent model over-fitting. We adopt weight decay by a factor of 0.1 whenever validation accuracy remains unchanged for 5 epochs. The batch size was set to be 100. All weights in the network are initialized using `He Normal` initialization.

4.4. Results

In order to evaluate our method quantitatively, we compute various measures and report the results below. Mean accuracy (mA) has been widely used in the attribute recognition literature [4]:

$$mA = \frac{1}{2N} \sum_{i=1}^{L} (TP_i/P_i + TN_i/N_i) \tag{1}$$

where L is the number of attributes, 35 for PETA and 51 for RAP dataset, respectively; P_i is number of positive examples, TP_i is the number of correctly predicted positive examples, N_i is the number of negative examples, and TN_i is the number of correctly predicted negatively examples. For each attribute, mA is average classification of the positive examples and the negative examples. Average over all attribute is then calculated as the final recognition rate.

The above measure treats each attribute independent of the other attributes. This might not necessarily be the case and an inter-attribute correlation might exist. Therefore, many researchers also report *example-based* evaluations [11]. These four metrics: accuracy, precision, recall, and F1 score are defined as:

$$Acc = \frac{1}{N} \sum_{N}^{i=1} \frac{|Y_i \bigcap f(x_i)|}{|Y_i \bigcup f(x_i)|} \tag{2}$$

$$Prec = \frac{1}{N} \sum_{N}^{i=1} \frac{|Y_i \bigcap f(x_i)|}{|f(x_i)|} \tag{3}$$

$$Rec = \frac{1}{N} \sum_{N}^{i=1} \frac{|Y_i \bigcap f(x_i)|}{|Y_i|} \tag{4}$$

$$F1 = \frac{2 \times Prec \times Rec}{Prec + Rec} \tag{5}$$

where, N is the total number of examples, Y_i is the ground truth label for the ith example, and $f(x)$ is the predicted label, respectively.

The test the method on two of the publicly available datasets and report the results on 35 attributes for the PETA dataset and 51 attributes for the RAP dataset. Table 2 shows a comparison of the proposed method with six current state of the art methods.

For the PETA dataset using **dscl_fc_based**, the obtained mA for our method is 80.28%. This is comparable to the other methods. Acc obtained from our method is 90.9%. This is higher than all the other methods that we compare with. The difference is almost 15%. The obtained results for the other measures (Pre, Rec and $F1$) is also comparable to the other methods. Fig. 4 depicts class-wise accuracies for the PETA dataset. As can be seen in the figure, the lowest accuracy

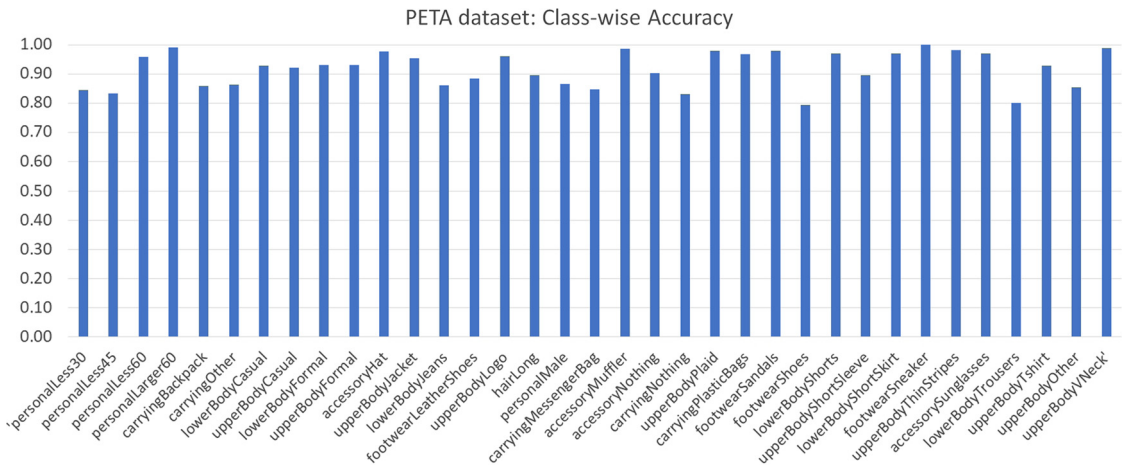

Fig. 4. Class-wise Accuracy - PETA dataset: the lowest accuracy is 79.27% for the class `footwearShoes`. The highest accuracy is for the class `footwearSneaker`.

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

	PETA [4]					RAP [11]				
	mA	Acc	Prec	Rec	F1	mA	Acc	Prec	Rec	F1
Chen et al. [17]	82.89	75.07	83.68	83.14	83.41	73.79	62.02	74.92	76.21	75.56
Sudowe et al. [41]	81.15	73.66	84.06	81.26	82.64	69.66	62.61	80.12	72.26	75.98
Liu et al. [15]	84.16	74.62	82.66	85.16	83.40	79.48	53.30	60.82	78.80	68.65
Sarfaraz et al. [24]	83.45	77.73	86.18	84.81	85.49	77.70	67.35	79.51	79.67	79.59
Li et al. [26]	81.77	76.13	84.92	83.24	84.07	76.12	65.39	77.33	78.79	78.05
Han et al. [39]	86.97	79.95	87.58	87.73	87.65	81.42	68.37	81.04	80.27	80.65
`dscl_fc_based` (ours)	80.28	**90.90**	85.10	78.75	81.60	**96.47**	**90.87**	80.92	73.47	73.97
`dscl_conv_based` (ours)	79.11	**87.40**	84.47	77.54	80.82	**95.45**	**91.10**	80.77	69.31	72.87

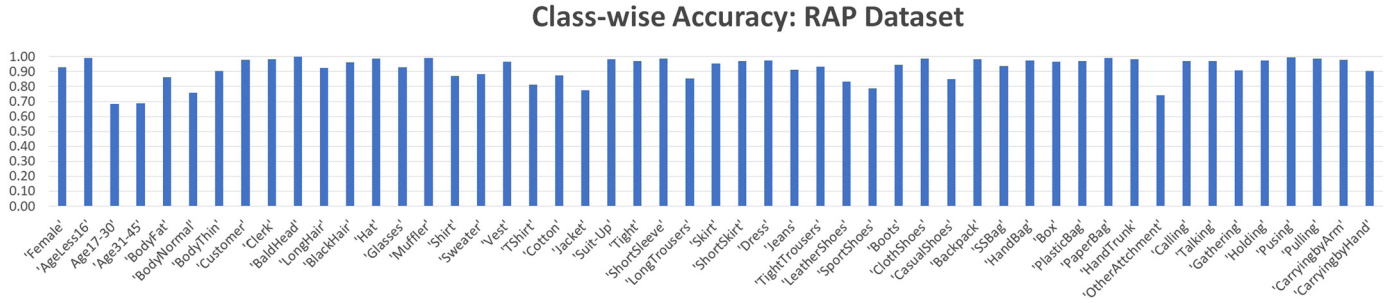

Fig. 5. Class-wise Accuracy - RAP dataset: the lowest accuracy is 68.3% for the class `Age17-30`. The highest accuracy is for the class `BaldHead`.

is that of the classes: `lowerBodyTrousers, and footwearShoes`. While the highest accuracy is that of the class `footwearSneakers`. These are some of the challenging classes, as the available information, in term of pixels, is very low for these scenarios. Some results from this dataset are shown in Fig. 6.

For the RAP dataset, similar to PETA dataset results, the trained network consistently out-performs other trained networks, but the difference is almost negligible. For mA, we obtained 96.87%, which is an improvement of approximately 16% over the state of the art. Similarly, we show a considerable improvement for the Acc, which is 90.87%, considerably higher than other methods. Similarly, the results obtained for $Prec$: 80.92%, Rec : 73.47% and $F1$: 73.97% is comparable to the other methods. Fig. 5 shows the class-wise accuracy for the RAP dataset. As can be seen in the figure, the highest accuracy is that of the class `BaldHead`. Two classes that have low accuracy are: `Age17-30, Age31-45`. This is obviously very difficult to judge even for human observers. Other low performing classes are: `Jacket, SportShoes, OtherAttachments`. The results for **`dscl_conc_based`** are also encouraging, although not better than **`dscl_fc_based`**. However, the number of parameters for this network is considerably lower, while not affecting heavily on the accuracy.

A demonstrated above, the proposed method performs comparable to other methods: mA and Acc is higher by a good margin than all other state of the art methods. For other measures, we perform equally well compared to the other methods. This is in addition to the fact that we are employing depthwise separable convolution layers, which reduces the number of parameters, resulting in efficient training. Moreover, we do not resort to data augmentation, perform part-based computations [11], learn attention-based network [26], compute pose estimation [17], or compute hand-crafted features [16].

These encouraging results demonstrate the use of separable convolution layers to the problem of pedestrian attribute recognition. We have trained the network on different color spaces and notice significant improvement. In addition, many of the existing methods fine-tune a pre-trained network (such as VGG19). In our view, existing pre-trained networks are not a good choice for the problem under consideration and considerably under-performed in our experiments. Our proposed method outperforms the standard deep learning methods on benchmark

Fig. 6. Some example recognition are shown here. We have a positive recognition of the left image to be wearing leather shoes and having age less than 30. A false positive is shown in the middle image as wearing shoes. Impressively, it is correctly recognized to be carrying an object even with a substantial occlusion. Image on the right is correctly recognized to be of age less than 30 and carrying a backpack, however, falsely it tagged as wearing formal upper body clothes.

public datasets. Our results are an improvement over the state of the art and demonstrate the practicality and the robustness of our approach.

5. Conclusion

In this paper, we have used the depthwise separable convolution layers, in contrast to the regular 2D convolution layers, and proposed a 3-branch Convolutional Neural Network (3bCNN). This has allowed us to train the network efficiently with reduced number of trainable parameters. We have trained the proposed network on YCrCb, L*a*b and HSV images. We have tested on two of the most challenging publicly available datasets (PETA and RAP). We have successfully demonstrated the effectiveness of the proposed method in comparison to the other state of the art works.

Declarations

Author contribution statement

Imran N. Junejo, Naveed Ahmed: Conceived and designed the experiments; Performed the experiments; 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.

Competing interest statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

[1] 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.

[2] 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.

[3] P. Viola, M. Jones, Robust real-time object detection, in: International Journal of Computer Vision (IJCV), vol. 57, 2001.

[4] 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.

[5] F. Chollet, Xception: deep learning with depthwise separable convolutions, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1800–1807.

[6] Z. Hu, et al., 3d separable convolutional neural network for dynamic hand gesture recognition, Neurocomputing 318 (2018) 151–161.

[7] F. Gonda, D. Wei, T. Parag, H. Pfister, Parallel separable 3d convolution for video and volumetric data understanding, in: BMVC, 2018.

[8] N. Hussein, E. Gavves, A.W.M. Smeulders, Timeception for complex action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, 2019, pp. 254–263.

[9] I.N. Junejo, A deep learning based multi-color space approach for pedestrian attribute recognition, in: Proceedings of the 2019 3rd International Conference on Graphics and Signal Processing, ICGSP '19, ACM, 2019, pp. 113–116.

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

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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 - Volume 37, ICML'15, 2015, pp. 448–456.

[20] P. Zhou, J. Han, G. Cheng, B. Zhang, Learning compact and discriminative stacked autoencoder for hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 57 (2019) 4823–4833.

[21] G. Cheng, C. Yang, X. Yao, L. Guo, J. Han, When deep learning meets metric learning: remote sensing image scene classification via learning discriminative cnns, IEEE Trans. Geosci. Remote Sens. 56 (2018) 2811–2821.

[22] G. Cheng, J. Han, P. Zhou, D. Xu, Learning rotation-invariant and Fisher discriminative convolutional neural networks for object detection, IEEE Trans. Image Process. 28 (2019) 265–278.

[23] G. Cheng, P. Zhou, J. Han, Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images, IEEE Trans. Geosci. Remote Sens. 54 (2016) 7405–7415.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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 - Volume 1, NIPS'14, MIT Press, Cambridge, MA, USA, 2014, pp. 487–495.

[30] 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.

[31] 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.

[32] 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.

[33] 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.
[34] X. Qian, et al., Pose-normalized image generation for person re-identification, in: The European Conference on Computer Vision (ECCV), 2018.
[35] P. Chikontwe, H.J. Lee, Deep multi-task network for learning person identity and attributes, IEEE Access 6 (2018) 60801–60811.
[36] 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.
[37] 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 33 (2019).
[38] X. Zhao, et al., Recurrent attention model for pedestrian attribute recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019) 9275–9282.
[39] K. Han, et al., Attribute aware pooling for pedestrian attribute recognition, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI'19, AAAI Press, 2019, pp. 2456–2462.
[40] F. Chollet, Keras, 2015.
[41] 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.