Pedestrian attribute recognition with part-based CNN and combined feature representations

Yiqiang Chen, Stefan Duffner, Andrei Stoian, Jean-Yves Dufour, Atilla Baskurt

Pedestrian attribute recognition with part-based CNN and combined feature representations

Yiqiang Chen[1], Stefan Duffner[1], Andrei Stoian[2], Jean-Yves Dufour[2] and Atilla Baskurt[1]
[1]*Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, France*
[2]*Thales Services, ThereSIS, Palaiseau, France*
{*yiqiang.chen, stefan.duffner, atilla.baskurt*}*@insa-lyon.fr,* {*jean-yves.dufour, andrei.stoian*}*@thalesgroup.com*

Keywords: Pedestrian attributes, Convolutional Neural Networks, Multi-label Classification

Abstract: In video surveillance, pedestrian attributes such as gender, clothing or hair types are useful cues to identify people. The main challenge in pedestrian attribute recognition is the large variation of visual appearance and location of attributes due to different poses and camera views. In this paper, we propose a neural network combining high-level learnt Convolutional Neural Network (CNN) features and low-level handcrafted features to address the problem of highly varying viewpoints. We first extract low-level robust Local Maximal Occurrence (LOMO) features and learn a body part-specific CNN to model attribute patterns related to different body parts. For small datasets which have few data, we propose a new learning strategy, where the CNN is pre-trained in a triplet structure on a person re-identification task and then fine-tuned on attribute recognition. Finally, we fuse the two feature representations to recognise pedestrian attributes. Our approach achieves state-of-the-art results on three public pedestrian attribute datasets.

1 INTRODUCTION

Pedestrian attributes are defined as semantic mid-level descriptions of people, such as gender, accessories, clothing *etc*. (see Fig. 1). Since biometric features like faces are often not visible or of too low resolution to be helpful in surveillance, pedestrian attributes could be considered as soft-biometrics and used in many surveillance applications like person detection (Tian et al., 2015), person retrieval (Vaquero et al., 2009), person identification (Layne et al., 2012) *etc*. A clear advantage of using attributes in this context is the possibility of querying a database of pedestrian images only by providing a semantic textual description (*i.e*. zero-short identification). Attributes have also been successfully used in object recognition (Duan et al., 2012), action recognition (Liu et al., 2011) and face recognition (Kumar et al., 2009).

The main challenges for pedestrian attribute recognition are the *large visual variation* and *large spatial shifts* due to the descriptions being on a high semantic level. For instance, the same type of clothes (e.g. shorts) can have very divers appearances. The large spatial shifts w.r.t. the detected pedestrian bounding boxes are caused by different body poses and camera views, and a finer body part detection or segmentation is challenging in surveillance-type videos. Furthermore, in realistic settings, illumination changes and occlusion make the problem even more challenging.

We propose a method to addresses these issues with the following contributions:

- High-level learnt features and low-level features are extracted and fused at a late training and processing stage to get a more robust feature representation. We will show that the two types of features are complementary and that combining them better models the divers appearances and locations of attributes.
- We propose to use a specific Convolutional Neural Network (CNN) architecture with 1D convolution layers operating on different parts of feature maps to model attribute patterns related to different body parts. In order to deal with large spatial shifts, we extract LOMO features (Liao et al., 2015) which have been specifically designed for viewpoint-invariant pedestrian re-identification.
- For small datasets, we propose to pre-train the deep neural network with re-identification data. This allows for a more effective attribute learning. We show that the knowledge learnt from the re-identification task can be transferred and help the attribute learning.

Figure 1: Some example images from pedestrian attribute datasets.

- Our method achieves state-of-the-art results on three public pedestrian attribute data sets: PETA, APiS and VIPeR.

2 RELATED WORK

Numerous approaches for pedestrian attribute recognition have been proposed in the past. Mid-level attributes were first used for human recognition by (Vaquero et al., 2009). The person image is parsed into regions, and each region is associated with a classifier based on Haar-like features and dominant colours. Then, the attribute information is used to index surveillance video streams. The approach proposed by (Layne et al., 2012) extracts a 2 784-dimensional low-level colour and texture feature vector for each image and trains an SVM for each attribute. The attributes are further used as a mid-level representation to aid person re-identification. (Zhu et al., 2013), in their work, introduced the pedestrian attribute database APiS. Their method determines the upper and lower body regions according to the average image and extracts colour and gradient histogram features (HSV, MB-LBP, HOG) in these two regions. Then, an Adaboost classifier is trained to recognise attributes. The drawback of these approaches is that all attributes are recognised independently, that is, the *relation* between attributes is not taken into account.

To overcome this limitation, (Zhu et al., 2014) proposed an interaction model, based on their Adaboost approach, learning an attribute interaction regressor. The final prediction is a weighted combination of the independent score and the interaction score. (Deng et al., 2014) constructed the pedestrian attribute dataset "PETA". Their approach uses a Markov Random Field (MRF) to model the relation between attributes. The attributes are recognised by exploiting the context of neighbouring images on the MRF-based graph. (Chen et al., 2017) uses a multi-label Multi-layer perceptron to classify the attributes in the same time.

With the recent success of Deep Learning for computer vision applications, methods based on Convolutional Neural Network (CNN) models have been proposed for pedestrian recognition. For example, (Li et al., 2015) fine-tuned the CaffeNet (similar to AlexNet (Krizhevsky et al., 2012)) trained on ImageNet to perform simple and multiple attribute recognition. (Zhu et al., 2015) proposed to divide the pedestrian images into 15 overlapping parts where each part connects to several CNN pipelines with several convolution and pooling layers.

Unlike these approaches that use either deep feature hierarchies or "hand-crafted" features, our method effectively fuses shift-invariant lower-level features with learnt higher-level features to build a combined representation that is more robust to the large intra-class variation which is inherent in attribute recognition. Recently, some deep features and "hand-crafted" features combination approaches have been also used in saliency detection(Li et al., 2017), face recognition(Lumini et al., 2016) and person re-identification(Wu et al., 2016).

We further address the large intra-class variation issue by a specific CNN architecture operating on different image regions related to the pedestrian body parts and using 1D horizontal convolutions on these part-based feature maps. We experimentally show that our system works well for both larger and smaller datasets thanks to a pre-training stage on the related task of pedestrian re-identification.

3 PROPOSED METHOD

Our approach takes as input a cropped colour (RGB) image of a pedestrian (resized to 128x48 pixels) and outputs a vector encoding the score for each attribute to recognise. The overall architecture of the proposed approach is shown in Fig. 2. The framework consists of two branches.

One branch extracts the viewpoint-invariant, hand-crafted Local Maximal Occurrence (LOMO) features. The extracted LOMO features are then projected into a linear subspace using Principal Component Analysis (PCA). The aim of this step is two-fold: first, to reduce the dimension of the LOMO feature vector removing potential redundancies, and second, to balance the contribution of CNN features and

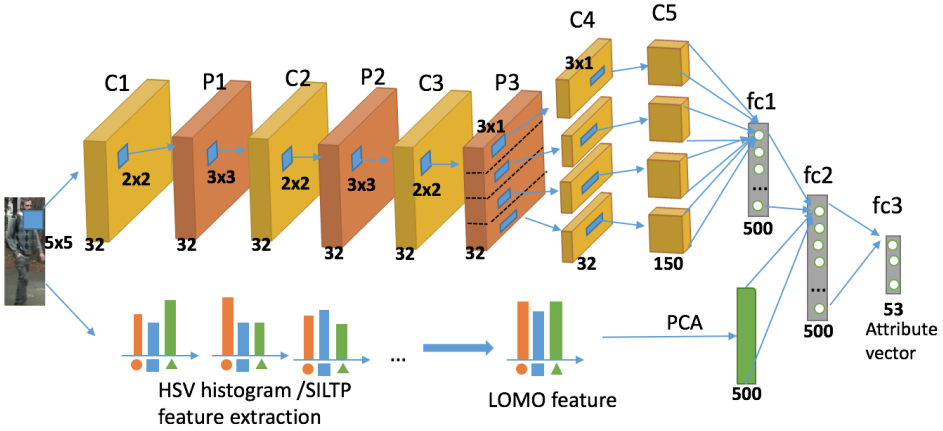

Figure 2: Overview of our pedestrian attribute recognition approach. Learnt features from a part-based CNN model are integrated with highly shift-invariant low-level LOMO features and used for multi-label classification.

LOMO features in the succeeding fusion that combines information represented in the two feature vectors.

The second branch is a Convolutional Neural Network extracting higher-level discriminative features by several succeeding convolution and pooling operations that become specific to different body parts at a given stage (P3) in order to account for the possible displacements of pedestrians.

To carry out this fusion, the output vectors of the two branches are concatenated and connected to two fully-connected layers (fc2+fc3) effectively performing the final attribute classification. We will explain these steps in more detail in the following.

3.1 LOMO feature extraction

Recently, pedestrian re-identification methods using LOMO feature (Liao et al., 2015) have achieved state-of-the-art performance, and here we apply these low-level features on the related task of attribute recognition in order to extract relevant cues from pedestrian images.

In the LOMO feature extraction method proposed by (Liao et al., 2015), the Retinex algorithm is integrated to produce a colour image that is consistent with human perception. To construct the features, two scales of Scale-Invariant Local Ternary Patterns (SILTP) (Liao et al., 2010) and an $8\times8\times8$-bin joint HSV histogram are extracted for an illumination-invariant texture and colour description. The sub-window size is 10×10, with an overlapping step of 5 pixels describing local patches in 128×48 images. Following the same procedure, features are extracted at 3 different image scales. For all sub-windows on the same image line, only the maximal value of the local occurrence of each pattern among these sub-windows is retained. In that way, the resulting feature vector achieves a large invariance to view point changes and, at the same time, captures local region characteristics of a person. We refer to (Liao et al., 2015) for more details.

In our approach, as illustrated at the bottom of Fig. 2, we project these extracted LOMO features of size 26 960 on a reduced linear subspace of dimension 500, in order to facilitate the later fusion and to remove most of the redundant information that is contained in these features. The projection matrix is learnt using PCA on the LOMO feature vectors computed on the training dataset.

3.2 Part-based CNN

In addition to the lower-level LOMO features which provide a higher invariance, we propose to extract deep feature hierarchies by a CNN model providing a higher level of abstraction and a larger discrimination power since the features are directly learnt from data.

As illustrated in Fig. 2, the CNN comprises three alternating convolution and pooling layers. The size of the first convolution (C1) is 5×5. The two following (C2, C3) are of size 3×3. The kernel size of max-pooling (P1-P3) is 2×2, and the number of channels of convolution and pooling layers is 32. The resulting feature maps in P3 are divided vertically into 4 equal parts which roughly correspond to the regions of head, upperbody, upperlegs and lowerlegs. The intuition behind this is that in pedestrian images the position of body parts varies much more horizontally than vertically due to the articulation of a walking person, for instance. Applying specific convolution filters on these different horizontal bands thus forces the CNN to extract features that are dedicated to dif-

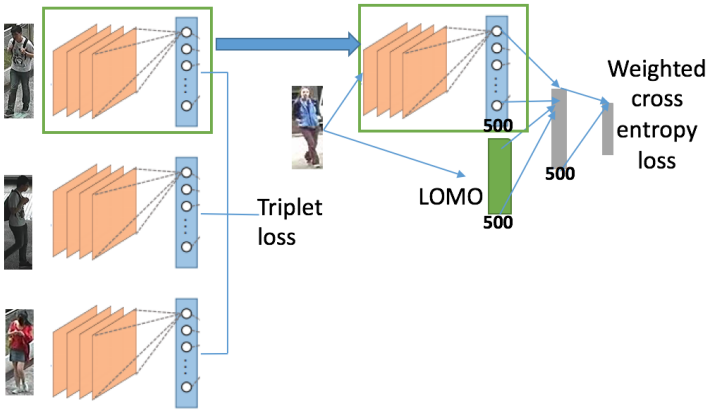

Figure 3: Illustration of the transfer learning from a re-identification task to attribute recognition. *Left:* the (shared) weights of the triplet CNN are pre-trained in a weakly supervised manner for pedestrian re-identification using the triplet loss function. *Right:* the CNN weights are integrated in our attribute recognition framework and the whole neural network is fine-tuned using the weighted cross-entropy loss.

ferent body parts and improves the overall learning and generalisation performance. For each part, similar to (Varior et al., 2016), we use two layers (C4, C5) with 1D horizontal convolutions of size 3×1 without zero-padding to reduce the feature maps to a single column. All the convolution layers in our model are followed by batch normalization and ReLU activation function (Krizhevsky et al., 2012). These 1D convolutions allow to extract high-level discriminative patterns for different horizontal stripes of the input image. In the last convolution layer, the number of channels is increased to 150, and these feature maps are given to a fully-connected layer (fc1) to generate an output vector of dimension 500.

Then this output vector and the projected LOMO feature vector are concatenated and processed by two further fully-connected layers (fc2, fc3) to perform the multi-label classification. This late fusion provides for a richer feature representation and robustness to viewpoint changes thanks to the shift-invariance property of LOMO and the body part modelling in our CNN architecture.

3.3 Training

To train the parameters of the proposed CNN, the weights are initialised at random and updated using stochastic gradient descent minimising the global loss function (*c.f.* Eq. 1) on the given training set. Since most attributes are not mutually exclusive, *i.e.* pedestrians can have several properties at the same time, the attribute recognition is a multi-label classification problem. Thus, the multi-label version of the sigmoid cross entropy is used as the overall loss function:

$$E = -\frac{1}{N}\sum_{i=1}^{N}\sum_{l=1}^{L}[w_l y_{il}\log(\sigma(x_{il}))+(1-y_{il})\log(1-\sigma(x_{il}))]\,, \tag{1}$$

$$\text{with}\quad \sigma(x) = \frac{1}{1+exp(-x)}\,,$$

where L is the number of labels (attributes), N is the number of training examples, and y_{il}, x_{il} are respectively the l^{th} label and classifier output for the i^{th} image. Usually, in the training set, the two classes are highly unbalanced. That is, for most attributes, the positive label (presence of an attribute) appears generally less frequently than the negative one (absence of an attribute). To handle this issue, we added a weight w to the loss function: $w = -log_2(p_l)$, where p_l is the positive proportion of attribute l in the dataset.

As we will show in our experiments, for smaller training dataset (like VIPeR), it is beneficial to pre-train the CNN with a (possibly larger) pedestrian re-identification dataset in a triplet architecture on the re-identification task, and then to fine-tune the pre-trained convolution layers on the actual attributes. Figure 3 illustrates this transfer learning approach.

Person re-identification consists in matching images of the same individuals across multiple camera views. In order to achieve this, we learn a distance function that has large values for images from different people and small values for images from the same person. A CNN with triplet architecture (Lefebvre and Garcia, 2013) can learn such a similarity function by effectively learning a projection on a (non-linear) subspace, where vectors from the same person are forced to be close and vectors from different persons are forced to be far. To this end, the network is presented with a triplet of pedestrian images composed

	Accuracy	Recall@FPR=0.1	AUC
LOMO (dim 500)	88.7	72.5	89.8
LOMO (dim 1000)	89.8	73.7	90.3
baseline	89.7	76.2	92.0
baseline + 2D conv	90.0	76.9	92.2
baseline + 1D conv	90.5	77.3	92.1
baseline + part-based 1D conv	90.8	78.7	92.3
baseline + 1D conv + LOMO (dim 1000)	91.5	79.4	91.7
baseline + part-based 1D conv + LOMO (dim 1000)	**91.7**	**81.3**	**93.0**

Table 1: Comparison of the 4 variants of our approach on PETA (in %).

of an anchor example a, a positive image p from the same person as the reference and a negative image n from a different person. The weights of the network for the three input images are shared. Let f(.) be the output of the CNN. Then the loss function is defined as:

$$E_{triplet} = -\frac{1}{N}\sum_{i=1}^{N}[max(\|f(a_i) - f(p_i)\|_2^2 - \|f(a_i) - f(n_i)\|_2^2 + m, 0)] \quad (2)$$

with m being a constant margin. The network gets updated when the negative image is closer than the positive image to the reference image. During training, for a given triplet, the loss function pushes the negative example away from the reference in the output feature space and pulls the positive example closer to it. Thus, by presenting many different triplet combinations, the network effectively learns a no-linear projection to a feature space that better represents the semantic similarity of pedestrians. The triplet architecture has been applied in object recognition (Wang et al., 2014), face recognition (Lefebvre and Garcia, 2013), person re-identification (Ding et al., 2015).

Unlike the identification task learning features that recognises the specific individuals, from the re-identification data, the triplet network learns informative features that distinguish individuals, and the semantic attributes that we want to recognise can be considered such identify features at a higher level. Therefore, this pre-learnt knowledge can be easily transferred to attribute recognition.

4 EXPERIMENTS

4.1 Datasets

We evaluated our approach on three public benchmarks: PETA, APiS and VIPeR (see Fig. 1).

The **PETA dataset** (Deng et al., 2014) is a large pedestrian attribute dataset which contains 19 000 images from several heterogeneous datasets. 61 binary attributes and 4 multi-class attributes are annotated. In our attribute recognition evaluation, we follow the experimental protocol of (Deng et al., 2014; Li et al., 2015): dividing the dataset randomly in three parts: 9 500 for training, 1 900 for validation and 7 600 for testing. Since different approaches (Deng et al., 2014; Li et al., 2015; Zhu et al., 2017) have been evaluated on different subsets of attributes, in our experiment we use the union of all these subsets, *i.e.* 53 attributes.

The **APiS dataset** (Zhu et al., 2013) contains 3 661 images collected from surveillance and natural scenarios. 11 binary attributes are annotated such as male/female, shirt, backpack, long/short hair. We followed the experimental setting of (Zhu et al., 2013). A 5-fold cross-validation is performed, and the final result is the average of the five tests.

The **VIPeR dataset** (Gray et al., 2007) contains 632 pedestrians in an outdoor environment, each having 2 images from 2 different view points. 21 attributes are annotated by (Layne et al., 2014). Each dataset is divided into two equal-size non-overlapping parts for training and testing (images from the same person are not separated). We repeat the process 10 times and report the average result.

During training, we perform data augmentation by randomly flipping and shifting the images slightly.

4.2 Parameter setting

All weights of the neural network are initialised from a Gaussian distribution with 0 mean and 0.01 standard deviation, and the biases are initialised to 0. The learning rate is set to 0.01. We used dropout (Srivastava et al., 2014) for the fully-connected layers with a rate of 0.6.

For tests on PETA, the fc1 layer, fc2 layer and PCA projected LOMO features are set to 1000 dimensions. The batch size is 100. For tests on APiS and VIPeR, fc1 fc2 layer sizes and PCA-projected LOMO feature size are 500 dimensions, and the batch size is 50.

The neural network is learned "from scratch" for tests on PETA and APiS. Since on VIPeR we have only 632 training images. The network is pre-trained with triplet loss on the CUHK03 dataset (Li et al.,

Attributes	Accuracy	Recall@FPR=0.1			AUC			
	ours	fusion(Zhu et al., 2013)	interact(Zhu et al., 2014)	ours	fusion(Zhu et al., 2013)	interact(Zhu et al., 2014)	DeepMar(Li et al., 2015)	ours
long jeans	93.5	89.9	89.2	**93.8**	96.1	96.2	96.5	**97.4**
long pants	94.2	78.7	80.6	**93.3**	92.5	93.9	**97.1**	**97.1**
M-S pants	93.7	76.7	85.1	**90.0**	92.4	92.8	95.5	**96.0**
shirt	88.4	68.2	**74.5**	65.5	83.9	83.9	**88.0**	87.3
skirt	95.6	58.3	61.3	**80.5**	90.0	**91.2**	91.0	90.5
T-shirt	79.6	56.2	56.5	**66.3**	85.4	85.5	**90.6**	88.7
gender	81.6	55.2	56.5	**65.1**	85.5	86.1	**90.0**	88.1
long hair	92.3	55.2	58.3	**68.9**	85.2	86.1	86.2	**88.1**
back bag	93.1	54.6	54.8	**61.2**	83.6	83.6	**86.6**	85.2
hand carrying	87.7	52.1	52.1	**60.6**	81.8	81.8	**84.3**	83.9
S-S bag	82.8	38.5	42.9	**54.0**	77.3	78.3	**83.7**	82.9
average	89.3	62.1	64.7	**72.7**	86.7	87.2	**90.0**	89.5

Table 2: Attribute recognition results on APiS (in %).

2014) which contains 13164 images of 1360 pedestrians. During training, the CNN part is fine-tuned with a lower learning rate (0.0005).

4.3 Evaluation

The test protocol on PETA (Deng et al., 2014) proposes to use the attribute classification accuracy. The APiS dataset's protocol (Zhu et al., 2013) uses the average recall at a False Positive Rate (FPR) of 0.1 and the Area Under Curve (AUC) of the average Receiver Operating Characteristics (ROC) curves as performance measures. As mentioned in (Zhu et al., 2017), accuracy is not sufficient to evaluate the classification performance on unbalanced attributes. In our experiment, we thus use all these three measures to evaluate our approach.

4.4 Results

We first evaluated the effectiveness of the 1D horizontal convolution layers, the body part division and the fusion of LOMO and CNN features using the PETA dataset. To show the effect of each contribution on the overall performance, we first implemented a baseline as a CNN with 3 consecutive convolution and max-pooling layers (C1-P3) and a multilayer perceptron using LOMO features of different PCA output dimensions as input. Then we implemented different variants of the proposed method: baseline with 2 layers of 3×3 convolution or 2 layers of 3×1 convolution, CNNs with and without body part division, and CNNs with and without LOMO feature fusion. Table 1 summarises the results.

We can conclude that the spatail invariance of the LOMO feature and the rich representation of the deep CNN features are compelentary and the fusion increases the overall recall and accuracy. 1D convolutions and dividing into body part also slightly improves the results. By performing all these, we obtain the highest overall accuracy, recall and AUC.

The comparison with the state of the art on PETA is shown in Table 3. In the literature, there are two evaluation settings for the PETA dataset with 35 and 45 attributes respectively. Table 3 shows the results on the 27 attributes that they have in common in order to compare all methods. We also display the average results for 35 and 45 attributes. Our method outperforms the state-of-the-art approach mlcnn by a margin of 3.4%, 14.3%, 6% points for the average accuracy recall and AUC on the 27 attributes and a margin of 3.5%, 15%, 6.1% points on the 45 attributes. We also outperform the DeepMar method by 9% points on accuracy. Moreover, our approach achieves a better score on almost all individual attributes.

The results on the APiS dataset are shown in Table 4. Our method outperforms the Adaboost approach with fusion features and interaction models by a margin of 6% and 2.3% points respectively for the recall at FPR=0.1 and AUC. Only for the AUC, DeepMar achieves a slightly better results (0.5% points) which could be explained by its pre-training on the large ImageNet dataset.

Finally, the results on the VIPeR dataset are shown in Table 2. Our approach achieves a 9.8% point improvement in accuracy and 4.1% points on recall at FPR=0.2 compared to the CNN-based state-of-the-art approach mlcnn-p. For most of the attributes, our method obtains a better score.

In summary, our approach outperforms the state of the art (including CNN-based methods) on two datasets and is on par with the best method on the third one. This demonstrates the robustness of the combined feature representation w.r.t. the high intra-class variation and the discriminative power of the proposed part-based CNN architecture.

Attributes	Accuracy Rate (%)				Recall@FPR=0.1		AUC	
	MRFr2(Deng et al., 2014)	DeepMar(Li et al., 2015)	mlcnn(Zhu et al., 2017)	ours	mlcnn(Zhu et al., 2017)	ours	mlcnn(Zhu et al., 2017)	ours
personalLess30	83.8	85.8	81.1	**86.0**	63.8	**80.8**	88.5	**93.8**
personalLess45	78.8	81.8	79.9	**84.7**	59.4	**74.9**	84.6	**91.9**
personalLess60	76.4	86.3	92.8	**95.4**	70.2	**83.0**	87.7	**92.8**
personalLarger60	89.0	94.8	97.6	**98.9**	90.7	**94.6**	94.9	**96.8**
carryingBackpack	67.2	82.6	84.3	**85.5**	58.4	**70.2**	85.2	**91.9**
carryingOther	68.0	77.3	80.9	**85.7**	46.9	**65.1**	77.7	**88.4**
lowerBodyCasual	71.3	84.9	90.5	**92.1**	56.2	**76.1**	87.5	**93.1**
upperBodyCasual	71.3	84.4	89.3	**91.2**	62.1	**74.2**	87.2	**92.5**
lowerBodyFormal	71.9	85.2	90.9	**93.3**	72.5	**82.8**	87.8	**92.7**
upperBodyFormal	70.0	85.1	91.1	**93.4**	70.5	**83.4**	87.6	**92.9**
accessoryHat	86.7	86.7	96.1	**97.5**	86.1	**89.9**	92.6	**95**
upperBodyJacket	67.9	79.2	92.3	**94.7**	53.4	**77.4**	81.0	**92.1**
lowerBodyJeans	76.0	85.7	83.1	**87.6**	67.6	**83.2**	87.7	**94.5**
footwearLeatherShoes	81.7	87.3	85.3	**90.2**	72.3	**87.8**	89.8	**95.7**
hairLong	72.8	88.9	88.1	**91.3**	76.5	**88.3**	90.6	**95.6**
personalMale	81.4	**89.9**	84.3	88.9	74.8	**87.0**	91.7	**95.8**
carryingMessengerBag	75.5	82.0	79.6	**84.5**	58.3	**70.7**	82.0	**89.8**
accessoryMuffler	91.3	96.1	97.2	**98.8**	88.4	**93.6**	94.5	**96.2**
accessoryNothing	80.0	85.8	86.1	**89.0**	52.6	**71.5**	86.1	**92.1**
carryingNothing	71.5	83.1	80.1	**84.5**	55.2	**71.8**	83.1	**91.3**
carryingPlasticBags	75.5	87.0	93.5	**96.6**	67.3	**83.6**	86.0	**92.2**
footwearShoes	73.6	80.0	75.8	**80.8**	52.8	**68.3**	81.6	**89.4**
upperBodyShortSleeve	71.6	87.5	88.1	**90.7**	69.2	**86.2**	89.2	**94.5**
footwearSneaker	69.3	78.7	81.8	**85.7**	52.0	**73.0**	83.2	**92.0**
lowerBodyTrousers	76.5	84.3	76.3	**83.4**	56.2	**75.2**	84.2	**92.0**
upperBodyTshirt	64.2	83.0	90.6	**93.3**	63.5	**82.7**	88.7	**92.8**
upperBodyOther	83.9	86.1	82.0	**86.2**	73.2	**80.8**	88.5	**93.5**
27 attributes average	75.8	85.4	86.6	**90.0**	65.6	**79.9**	87.0	**93.0**
35 attributes in (Deng et al., 2014; Li et al., 2015) average	71.1	82.6		**91.7**		78.9		**92.0**
45 attributes in (Zhu et al., 2017) average			87.2	**90.7**	67.3	**82.3**	87.7	**93.8**
53 attributes average				91.7		81.3		93.0

Table 3: Attribute recognition results on PETA (in %).

5 CONCLUSION

In this paper, a pedestrian attribute classification approach based on deep learning has been proposed. Our approach applies 1D convolutions on part-based feature map and fuses low-level LOMO features and high-level learnt CNN features to construct an effective classifier that is robust to large view point and pose variations. We proved that the learned CNN features and the hand craft LOMO features are complementary and the fusion improves the attribute recognition results. We also showed that pre-training the CNN model on person re-identification can assist attribute learning for small datasets. Finally, in our experiments on three public benchmarks, the proposed approach showed superior performance compared to the state of the art.

ACKNOWLEDGEMENTS

This work was supported by the Group Image Mining (GIM) which joins researchers of LIRIS Lab. and THALES Group in Computer Vision and Data Mining. We thank NVIDIA Corporation for their generous GPU donation to carry out this research.

Attributes	Accuracy			Recall@FPR=0.2			AUC
	svm(Layne et al., 2012)	mlcnn-p(Zhu et al., 2015)	ours	svm(Layne et al., 2012)	mlcnn-p(Zhu et al., 2015)	ours	ours
redshirt	85.5	91.9	**94.4**	88.4	88.9	**95.9**	95.2
blueshirt	73.0	69.1	**91.5**	60.8	70.8	**75.5**	83.1
lightshirt	83.7	83.0	**84.4**	87.8	85.3	**88.2**	91.7
darkshirt	**84.2**	82.3	83.3	**87.5**	85.8	86.1	90.9
greenshirt	71.4	75.9	**96.2**	54.3	69.4	**84.6**	88.7
nocoat	70.6	71.3	**74.2**	59.3	57.2	**65.4**	80.4
notlightdarkjean	70.3	90.7	**96.7**	57.2	78.6	**80.0**	86.0
darkbottoms	75.7	78.4	**78.9**	70.2	**76.2**	74.9	85.7
lightbottoms	74.7	76.4	**76.5**	69.5	**73.3**	72.3	83.6
hassatchel	47.8	57.8	**70.9**	22.0	31.7	**39.1**	64.8
barelegs	75.6	84.1	**92.2**	68.7	85.4	**92.2**	92.8
shorts	70.4	81.7	**92.3**	59.8	82.9	**87.3**	88.6
jeans	76.4	77.5	**80.6**	72.7	74.7	**81.7**	87.6
male	66.5	69.6	**74.7**	48.2	57.2	**67.9**	82.1
skirt	63.6	78.1	**94.3**	40.7	60.7	**61.3**	72.8
patterned	46.9	57.9	**90**	26.3	41.0	**49.9**	68.1
midhair	64.1	**76.1**	75.2	43.0	**63.5**	54.1	73.1
darkhair	63.9	**73.1**	67.5	39.6	**58.4**	49.7	71.9
hashandbagcarrierbag	45.3	42.0	**90.9**	17.4	18.5	**27.5**	55.1
hasbackpack	67.5	64.9	**72.7**	47.9	49.9	**57.4**	76.3
average	68.9	74.1	**83.9**	56.1	65.5	**69.6**	80.9

Table 4: Attribute recognition results on VIPeR (in %).

REFERENCES

Chen, Y., Duffner, S., Stoian, A., Dufour, J.-Y., and Baskurt, A. (2017). Triplet cnn and pedestrian attribute recognition for improved person re-identification. In *Proceedings of the IEEE International Conference on Advanced Video and Singal based surveillance*.

Deng, Y., Luo, P., Loy, C. C., and Tang, X. (2014). Pedestrian attribute recognition at far distance. In *Proc. of the ACM international conference on Multimedia*, pages 789–792.

Ding, S., Lin, L., Wang, G., and Chao, H. (2015). Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003.

Duan, K., Parikh, D., Crandall, D., and Grauman, K. (2012). Discovering localized attributes for fine-grained recognition. In *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3474–3481.

Gray, D., Brennan, S., and Tao, H. (2007). Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. of International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proc. of Advances in neural information processing systems (NIPS)*, pages 1097–1105.

Kumar, N., Berg, A. C., Belhumeur, P. N., and Nayar, S. K. (2009). Attribute and simile classifiers for face verification. In *Proc. of the International Conference on Computer Vision (ICCV)*, pages 365–372.

Layne, R., Hospedales, T. M., and Gong, S. (2014). Attributes-based re-identification. In *Person Re-Identification*, pages 93–117. Springer.

Layne, R., Hospedales, T. M., Gong, S., and Mary, Q. (2012). Person re-identification by attributes. In *Proc. of the British Machine Vision Conference (BMVC)*, page 8.

Lefebvre, G. and Garcia, C. (2013). Learning a bag of features based nonlinear metric for facial similarity. In *Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on*, pages 238–243. IEEE.

Li, D., Chen, X., and Huang, K. (2015). Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. *Proc. of the Asian Conference on Pattern Recognition (ACPR)*.

Li, H., Chen, J., Lu, H., and Chi, Z. (2017). Cnn for saliency detection with low-level feature integration. *Neurocomputing*, 226:212–220.

Li, W., Zhao, R., Xiao, T., and Wang, X. (2014). Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159.

Liao, S., Hu, Y., Zhu, X., and Li, S. Z. (2015). Person re-identification by local maximal occurrence representation and metric learning. In *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2197–2206.

Liao, S., Zhao, G., Kellokumpu, V., Pietikäinen, M., and Li, S. Z. (2010). Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In *Proc. of the IEEE International*

Conference on Computer Vision and Pattern Recognition (CVPR), pages 1301–1306.

Liu, J., Kuipers, B., and Savarese, S. (2011). Recognizing human actions by attributes. In *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3337–3344.

Lumini, A., Nanni, L., and Ghidoni, S. (2016). Deep featrues combined with hand-crafted features for face recognition. *International Journal of Computer Research*, 23(2):123.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958.

Tian, Y., Luo, P., Wang, X., and Tang, X. (2015). Pedestrian detection aided by deep learning semantic tasks. In *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5079–5087.

Vaquero, D. A., Feris, R. S., Tran, D., Brown, L., Hampapur, A., and Turk, M. (2009). Attribute-based people search in surveillance environments. In *Proc. of Workshop on Applications of Computer Vision (WACV)*, pages 1–8.

Varior, R. R., Haloi, M., and Wang, G. (2016). Gated siamese convolutional neural network architecture for human re-identification. In *Proc. of the IEEE International Conference on European Conference on Computer Vision (ECCV)*, pages 791–808.

Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., and Wu, Y. (2014). Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1393.

Wu, S., Chen, Y.-C., Li, X., Wu, A.-C., You, J.-J., and Zheng, W.-S. (2016). An enhanced deep feature representation for person re-identification. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8. IEEE.

Zhu, J., Liao, S., Lei, Z., and Li, S. Z. (2014). Improve pedestrian attribute classification by weighted interactions from other attributes. In *Asian Conference on Computer Vision*, pages 545–557.

Zhu, J., Liao, S., Lei, Z., and Li, S. Z. (2017). Multi-label convolutional neural network based pedestrian attribute classification. *Image and Vision Computing*, 58:224–229.

Zhu, J., Liao, S., Lei, Z., Yi, D., and Li, S. Z. (2013). Pedestrian attribute classification in surveillance: Database and evaluation. In *Proc. of the International Conference on Computer Vision (ICCV) Workshops*, pages 331–338.

Zhu, J., Liao, S., Yi, D., Lei, Z., and Li, S. Z. (2015). Multi-label cnn based pedestrian attribute learning for soft biometrics. In *Proc. of the International Conference on Biometrics(ICB)*, pages 535–540.