HAL
open science

Deep and Low-level Feature based Attribute Learning for Person Re-identification

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

To cite this version:

Yiqiang Chen, Stefan Duffner, Andrei Stoian, Jean-Yves Dufour, Atilla Baskurt. Deep and Low-level Feature based Attribute Learning for Person Re-identification. Image and Vision Computing, 2018, 79, pp.25 - 34. 10.1016/j.imavis.2018.09.001 . hal-01895367

HAL Id: hal-01895367

https://hal.science/hal-01895367v1

Submitted on 15 Oct 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Deep and Low-level Feature based Attribute Learning for Person Re-identification

Yiqiang Chen[1], Stefan Duffner[1,], Andrei Stoian[1,], Jean-Yves Dufour[1,], Atilla Baskurt[1,]

[a] *Universiteé de Lyon, CNRS*
INSA-Lyon, LIRIS, UMR5205, France
[b] *Thales Services, ThereSIS, Palaiseau, France*

Abstract

In video surveillance, pedestrian attributes are defined as semantic descriptors like gender, clothing or accessories. In this paper, we propose a CNN-based pedestrian attribute-assisted person re-identification framework. First we perform the attribute learning by a part-specific CNN to model attribute patterns related to different body parts and fuse them with low-level robust Local Maximal Occurrence (LOMO) features to address the problem of the large variation of visual appearance and location of attributes due to different body poses and camera views. Our experiments on three public benchmarks show that the proposed method improves the state of the art on attribute recognition. Then we merge the learned attribute CNN embedding with another identification CNN embedding in a triplet structure to perform the person re-identification task. Both CNNs are pre-trained in a supervised way on attributes and person identities respectively, and then continue the training with a combined architecture for re-identification. We experimentally show that this fusion of "identity and attributes features" improves the overall re-identification.

Keywords: Person re-identification; soft-biometrics; pedestrian attributes; convolutional neural network

1. Introduction

With the increase of the number of surveillance cameras and public security demand, the video surveillance analysis tasks including object detection, object tracking or abnormal event detection become important research topics developing methods to effectively process these large amounts of data. In this paper, we focus on person re-identification by pedestrian attributes.

Recently, visual attributes recieved a lot of attention and have been used for object recognition [**?**], action recognition [**?**] , face recognition [**?**] etc. Pedestrian attributes are defined as semantic mid-level descriptions of persons, such as gender, accessories, clothing and so on. The advantage of attributes is that they are more robust to visual changes related to the viewing angle, body pose or lighting for instance, and that they can be used for zero-shot identification (querying by an attribute-based description instead of an image). 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 provide helpful information for many surveillance applications like person detection [**?**], person retrieval [**?**], or abnormal event detection. For example, a description like "a male in black shirt with a back bag" can be effectively used in person retrieval applications.

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.

Person re-identification consists in matching a query person among a large set of people detected in multiple non overlapping camera views. The challenges for automatic re-identification approaches from images are essentially the same

as for pedestrian attribute recognition i.e. a large variation in viewpoint, pose, illumination and background.

This paper is an extension of our preliminary work on attribute learning [**?**]. In this paper, we present a CNN-based pedestrian attribute assisted person re-identification framework. In the first step, the attribute learning is performed. In order to deal with the large spatial shift of attributes, we propose to use a specific Convolution Neural Network (CNN) architecture with 1D convolution layers operating on several horizontal parts of the input feature maps to learn different feature representations and model the displacements of different body parts. For an even larger spatial invariance, our approach extracts LOMO features, which have been specifically designed for viewpoint-invariant pedestrian re-identification. These low-level handcrafted features are fused with the high-level learned CNN features at a late training and processing stage to get a more robust feature representation modelling the diverse appearances of attributes. Our experiments show that the proposed method improves the state of the art on pedestrian attribute recognition on three public benchmarks.

In the second step, the learned attribute embedding is used for person re-identification. The framework fuses two neural networks. One is our attribute recognition network pre-trained with attribute labels, the other is a CNN pre-trained with person identity labels. Then we integrate these two neural networks into a triplet architecture to learn the optimal fusion parameters. To this end, an improved triplet loss with hard example selection is used. We experimentally show that the fusion leads to a better re-identification performance, and our approach achieves state-of-the-art results.

2. Related work

2.1. Pedestrian Attribute Recognition

There are many approaches for attribute recognition like clothing description and human attributes. Particularly, pedestrian attribute recognition from surveillance videos has attracted much attention, where attributes correspond

to mid-level semantic descriptions(gender, hair type, closthing, accessories etc). In the approach proposed by Vaquero et al. [?], the person image is parsed into regions, and each region is associated with a classifier based on Haar-like features and dominant colors. The attribute information is then used to index surveillance video streams. Layne et al. [?] annotated 15 attributes on the VIPeR dataset and proposed an approach to extract a 2 784-dimensional low-level color and texture feature vector for each image and to train an SVM for each attribute. Zhu et al. [?], 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 color and texture features (HSV, MB-LBP, HOG) in these two regions. Then, an Adaboost classifier is trained on these features to recognize attributes. The drawback of these approaches is that all attributes are treated independently. That is, the relation between different attributes is not taken into account.

Some later works try to overcome this limitation. Zhu et al. [?] 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. [?] constructed the pedestrian attribute dataset Peta and their approach uses a Markov Random Field (MRF) to model the relation between attributes. The attributes are recognized by exploiting the context of neighbouring images on the MRF-based graph.

Some Convolution Neural Network (CNN) models have been proposed for pedestrian recognition. For example, Li et al. [?] fine-tuned the CaffeNet (similar to AlexNet) trained on ImageNet to perform simple and multiple attribute recognition. Similarly, Sudowe et al [?] proposed the Attribute Convolutional Net (ACN) which add custom layers on Alexnet to jointly learn attributes. Zhu et al. [?] proposed to divide the pedestrian images into 15 overlapping parts where each part connects to several CNN pipelines with several convolution and pooling layers. They further pre-define connections between the parts and the attributes in the fully-connected layers to deal with the shift problem. Later

they proposed an improved version [?], where a fully-connected layer connect to all attributes in stead of using manually defined connections.

Recently, some deep features and hand-crafted features combination approaches have been proposed in different tasks like saliency detection [?], face recognition [?] and person re-identification [?]. These approaches implement deep neural network framework embedded with low-level features. In our work, we exploit this handcrafted and deep feature combination in the attribute recognition context. Our method effectively fuses shift-invariant lower-level features with learned higher-level features to build a combined representation that is more robust to the large intra-class variation which is inherent in attribute recognition.

We implemented also 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. The detected attributes are further used to assist the person re-identification task.

2.2. Person re-identification

Existing person re-identification approaches generally build a robust feature representation or learn a distance metric. The features used for re-identification are mainly variants of color histograms [? ?], Local Binary Patterns (LBP) [? ?] or Gabor features [?]. For example, Farenzena et al. [?] partitioned the human body into meaningful parts exploiting asymmetry and symmetry principles. On each part, the weighted color histogram, the maximally stable color regions and the recurrent high-structured patches are computed. Ma et al. [?] use local descriptors based on color and gradient information and encode them using high-dimensional Fisher vectors. Mignon et al. [?] used a feature vector composed of a mixture of color (RGB, HSV and YCbCr) and texture (LBP) from six horizontal stripe regions. Liao et al. [?] analyse the horizontal occurrence of local features and maximize this occurrence to

improve the robustness of features. The main metric learning methods include Mahalanobis metrics like KISSME [?], Local Fisher discriminant Analysis (LFDA) [?], Marginal Fisher Analysis(MFA) [?] and Cross-view Quadratic Discriminant Analysis (XQDA) [?].Other than Mahalanobis metrics, SVM [?] and boosting[?] have also been proposed for re-identification metric learning.

Recently some CNN models have been proposed for person re-identification. Yi et al. [?] first applied a Siamese network to person re-identification. The person image is divided into three parts, each associated with a CNN. To handle geometric problems, DeepReId [?] implements a novel architecture where a patch matching layer models the displacement of body parts. An improved CNN architecture proposed by Amed et al.[?] computes the cross-input neighborhood difference features. Cheng et al. [?] introduced a variant of the triplet loss function and a CNN network processing parts and the entire body. In [?], a gate layer is integrated in a Siamese CNN to capture effective subtle patterns in the feature map. Shi et al. [?] proposed a moderate positive mining method which improve the learning efficiency in a Siamese-like CNN framework. liu et al. [?] proposed a soft attention based Siamese neural network.

Some previous works have used pedestrian attributes to assist with the re-identification task. Based on an attribute recognition SVM approach, Layne et al. [?] first proposed to use attributes as a mid-level representation for improving person re-identification. The final distance between two pedestrian images is computed as a weighted sum of low-level feature distance and attribute distance. Li et al. [?] embeds middle-level clothing attributes via a latent SVM framework for more robust person re-identification. The approach introduced by Khamis et al. [?] learns a discriminative projection to a joint appearance-attribute subspace in order to leverage the interaction between attributes and appearance for matching.

Some other related methods also use attributes to perform person re-identification with CNNs, but the ways of integrating these attributes are different, for example using a weighted combination, simple concatenation, multi-task learning or attribute pre-training. Zhu et al. [?] recognize the attributes with deep

neural networks then calculate a pedestrian distance by weighting the attribute distance and a low-level feature-based person appearance distance. McLaughlin et al. [**?**] propose to perform person re-identification and attribute recognition in a multi-task learning. this proposed loss function is a weighted sum of the attribute and identification classification loss as well as a Siamese loss. And they show a re-identification performance improvement by multitask joint learning. Matsukawa et al. [**?**] propose to finetune the well-known Alexnet with attribute combination labels to increase the discriminative power. Further they concatenated the CNN embedding directly with LOMO features and used the metric learning method XQDA [**?**] to learn a feature space. Su et al. [**?**] proposed a three-stage procedure that pre-trains a CNN with attribute labels of an independent dataset, then fine-tunes the network with identity labels and finally fine-tunes the network with the learned attribute feature embedding on the combined dataset. The main difference of these approaches to ours is the way of making use of attributes to assist in the re-identification task. In summary, two CNN embeddings are learned based on attribute and identity annotation. Then, an improved triplet loss is used to learn the fusion. We will experimentally show the performance improvement brought by this fusion achieving state-of-art results on a public person re-identification benchmark.

3. Proposed Methods

In this section, we first describe our attribute recognition method and further introduce the attribute and identity-based re-identification framework.

3.1. CNN based pedestrian attribute recognition

3.1.1. Overall procedure

The architecture of the proposed attribute recognition approach is shown in Fig. **??**. The framework consists of two branches. One branch is a Convolutional Neural Network extracting higher-level discriminative features by several succeeding convolution and pooling operations that become specific to

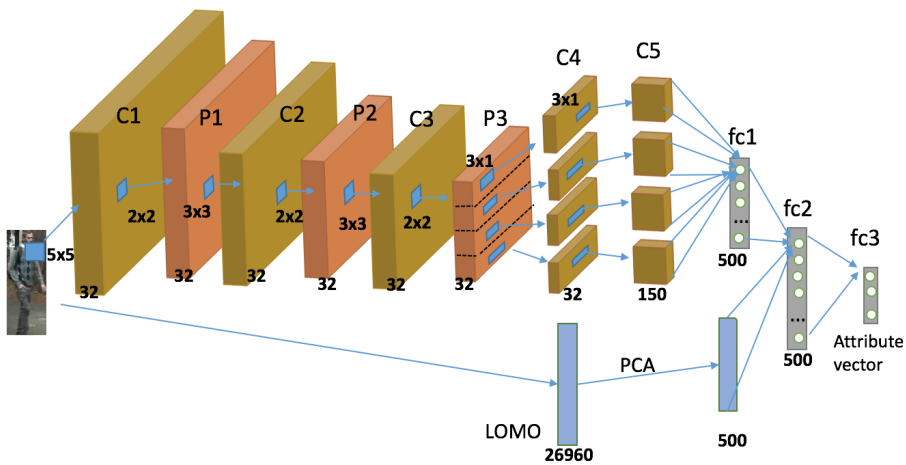

Figure 1: Overview of attribute recognition approach

different body parts at a given stage (P3) in order to account for the possible displacements of pedestrians due to pose variations. Another branch extracts the viewpoint-invariant Local Maximal Occurrence (LOMO) features, a robust visual feature representation that has been specifically designed for viewpoint-invariant pedestrian attribute recognition and achieving state-of-the-art results [**?**] (cf. Section **??**). 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 LOMO features in the succeeding fusion that combines information represented in the two feature vectors.

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

3.1.2. Part-based CNN

We first 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 learned 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 respectively. The resulting feature maps in the last pooling layer are divided vertically into 4 equal parts roughly corresponding to the regions of head, upper body, upper legs and lower legs. For each part, similar to [**?**], we use two layers (C4, C5) with 1D horizontal convolutions of size 3×1 without zero-padding reducing the feature maps to single column vectors. 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. All the convolution layers in our model are followed by batch normalization and ReLU activation functions.

3.1.3. LOMO extraction

Recently, pedestrian re-identification methods using LOMO features proposed by [**?**] 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 and to complement the CNN features by providing a higher viewpoint invariance. In the LOMO feature extraction method, 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) [**?**] and an 8×8×8-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, fea-

tures 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 [**?**] for more details. In our approach, as illustrated in the bottom of Fig. 2, we perform a dimensionality reduction projecting the extracted LOMO features of size 26 960 on a linear subspace of dimension 500, in order to facilitate the later fusion. The projection matrix is computed using PCA on the LOMO feature vectors computed on the training dataset.

3.1.4. 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 (Eq. **??**) 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:

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

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

where L is the number of labels (attributes), N is the number of training examples, and y_{il}, x_{il} are respectively the lth label and classifier output for the ith image. Usually, in the training set, the two classes are highly unbalanced. That is, for most attributes, the positive label appears generally less frequently than the negative one. 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

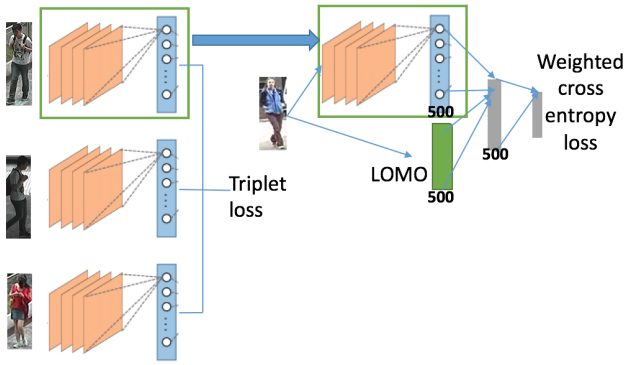

Figure 2: Illustration of the transfer learning from a re-identification task to attribute recognition. *Left:* the (shared) weights of the triplet CNN are pre-trained in a weakly supervised manner for pedestrian re-identification using the triplet loss function. *Right:* the CNN weights are integrated in our attribute recognition framework and the whole neural network is fine-tuned using the weighted cross-entropy loss.

re-identification dataset in a triplet architecture on the re-identification task. Since pedestrian attribute recognition and re-identification are two similar tasks, the visual features learned from re-identification can be useful for recognising attributes. Thus, after this pre-training, we transfer the re-identification knowledge to attribute recognition by fine-tuning the pre-trained convolution layers on the actual small attribute datasets. Figure **??** 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 need to 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 [**?**] 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 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:

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

with m being a constant margin. The network gets updated when the negative image is nearer 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.

From the re-identification data, the 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-learned knowledge can be effectively transferred to this problem. In the next section, we will also show that, inversely, attribute information can support the re-identification task.

3.2. Person re-identification

We propose a new CNN-based approach for pedestrian re-identification, that effectively combines automatically learned visual features with semantic attributes. To this end, we make use of our attribute recognition neural network presented in Section **??**. Attributes are important cues for a human to identify persons by appearance. The attribute learning consists in representing data instances by projecting them onto a basis set defined by domain-specific axes which are semantically meaningful to humans. Compared to features that are directly learned form appearance, semantic attributes are more consistent for the same person and are more robust to the different variations. Since the attribute information adds additional constraints to person identity consistency, i.e. appearance consistency and attribute consistency, the combination with

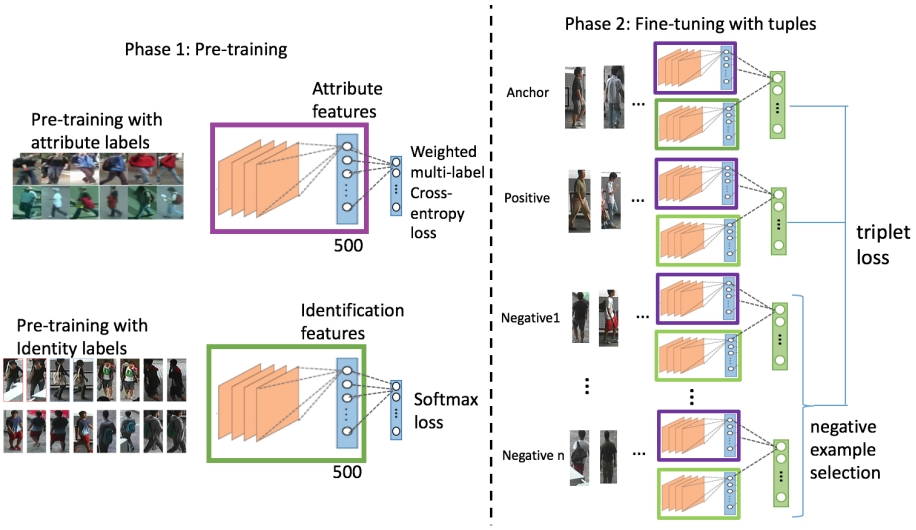

Figure 3: Overall architecure of our re-identification method

the attribute CNN embedding which encodes these attribute constraints improves the person re-identification performance, as we will show experimentally in Section **??**.

The overall framework is shown in figure **??**. The framework is composed of two neural networks that are pre-trained. The first is a CNN that is trained in a supervised way to classify identities on a separate training set. Then we remove the output classification layers of the network and keep the other parts of the network which are related to feature selection. The second part is our attribute recognition network that is trained as described in section **??**. After training, we also remove the output layers and keep all the other layers up to the first fully-connected layer (fc1).

The output vectors from the hidden layers of the two CNNs represent high-level features related to attributes and pedestrian identities respectively. As we will show experimentally, the information extracted by the two CNNs is complementary, thus using both leads to an overall performance improvement. In order to combine the extracted features effectively, we propose to integrate both output vectors in a new neural network that automatically learns these

layer	type	filter size	padding	output size
C1	Convolution	5×5	yes	128×48×32
P1	Max-Pooling	2×2	-	64×24×32
C2	Convolution	3×3	yes	64×24×32
P2	Max-Pooling	2×2	-	32×12×32
C3	Convolution	3×3	yes	32×12×64
P3	Max-Pooling	2×2	-	16×6×64
C4	Convolution	3×3	yes	16×6×128
P5	Max-Pooling	2×2	-	8×3×128
C5	Convolution	3×1	no	8×1×400
fc1	Fully-connected	-	-	500
fc2	Fully-connected	-	-	N

Table 1: Identification network parameters.

fusion parameters on the re-identification task in a triplet architecture. this leads to a fully neural architecture that can be trained and fine-tuned as a whole to maximise the re-identification performance. We will explain these steps in more detail in the following.

3.2.1. Supervised identification CNN

The identification network consists of 5 convolutional layers and 4 max-pooling layers. The details are presented in Table. **??**. The first convolutional layer has a kernel size of 5x5 and the following 3 convolutional layers have a kernel size of 3×3. The last one has a kernel size of 3×1 without zero-padding increasing the number of channels but reducing their size to a single column. At the end, there are two fully-connected layers. All max pooling layers all have a kernel size of 2×2. The output of the network is a vector of N dimensions with N being the number of identities in the training set. Batch normalization and ReLU activation function are applied after the convolutional layers and fully connected layers. As mentioned earlier, this CNN is pre-trained in a supervised way, using images and identity labels from a separate training dataset. To this end, we minimise the following softmax corss-entropy loss on the given

classification task:

$$P(y_j = 1|x) = \frac{e^{W_j^T x + b_j}}{\sum_{k=1}^{N} e^{W_k^T x + b_k}} \tag{3}$$

$$E_{identification} = -\sum_{k=1}^{N} y_k log(P(y_k = 1|x)) \tag{4}$$

Where N is the number of identities. y is the one-hot identity label. x is the input of the last fully connected layer. W and b are weight and bias term of the last fully connected layer. $P(y_j = 1|x)$ is the probability predicted that the input x corresponds to identity j. The intuition of this is that this learned feature representation can be used for learning similarities between arbitrary pedestrian images and thus be transferred to the task of re-identification.

3.2.2. Fusion by Triplet architecture

The pre-trained attribute CNN and identification CNN are combined and trained in a triplet architecture similar to the one explained in section **??**. Here, we propose to use an improved triplet loss with hard example selection to learn the optimal fusion of the two types of features. The fc1 layer of the attribute network and the fc1 layer of the identification network are normalized and concatenated, and another fully-connected layer which allows to merge attribute and identification features. Unlike with classic triplet loss, a (n+2)-tuple of images instead of a triplet is projected into the feature space. The tuple includes one anchor image, one positive image and n negative images. Training enforces that the projection of the positive example is placed closer to the anchor than the projection of the closest negative example among n negative examples. This constraint is defined as following:

$$min(\|f(a) - f(n^j)\|_2^2) - \|f(a) - f(p)\|_2^2 > m \tag{5}$$

Similarly to the distance learning approach "Top-push" proposed by [**?**] , hard example mining in [**?** **?**] or moderate positive example mining in [**?**], the idea is finding the appropriate example to update the model. The negative example

that is closest to the anchor is considered the hardest example and having the highest potential for improvement. The network is thus updated efficiently by pushing the hardest example further away from the anchor. The intuition is that if the positive example is ranked in front of the hardest negative example then the positive example is ranked first, which is our goal. In classic triplet loss, a big part of triplets does not violate the triplet constraint (c.f Eq **??**). These triplets are useless for learning. The selection among n negative examples reduces the number of unused training data and can make the training more efficient.

To further enhance the loss function, as an extention of [**?**], we add a term including distance between the anchor example and the positive example. The loss function is defined as follows:

$$E_{min-triplet} = -\frac{1}{N}\sum_{i=1}^{N}[max(\|f(a_i) - f(p_i)\|_2^2 \\ - min(\|f(a_i^j) - f(n_i^j)\|_2^2) + m, 0) + \alpha\|f(a_i) - f(p_i)\|_2] \quad (6)$$

The first part of the loss is a comparison of two distances which defines a relative relationship in the feature space. The second part corresponds to an absolute distance in feature space. Combining these two constraints leads to a more efficient learning of the resulting manifold that better represents the semantic similarities.

Using this loss function, we train the additional fully-connected layer for the fusion, and, at the same time we fine-tune the other part of the network, i.e. the weights are updated at a lower rate. Since pedestrian attributes are difficult to annotate, especially for large re-identification dataset. Unlike other approaches[**?** **?** **?** **?**], the advantage of our method is that the attributes do not need to be annotated on the re-identification dataset. We can make use of a separate attributes dataset with annotated attributes and transfer this information to a re-identification dataset by fine-tuning.

Figure 4: Some example images from pedestrian attribute datasets.

4. Experiments

In this section, the proposed attribute recognition methods are evaluated on the VIPeR [?] dataset with the annotation from [?], PETA dataset[?] and the APiS dataset [?] (see Fig. 4). Finally, we test our proposed CNN architecture for person re-identification on the CUHK03 dataset [?].

4.1. attribute recognition experiments

4.1.1. Datasets

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

The **PETA dataset**[?] 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 [? ?]: dividing the dataset randomly in three parts: 9 500 for training, 1 900 for validation and 7 600 for testing. Since different approaches[? ? ?] 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** [**?**] 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 [**?**]. A 5-fold cross-validation is performed, and the final result is the average of the five tests.

The **VIPeR dataset** [**?**] contains 632 pedestrians in an outdoor environment, each having 2 images from 2 different view points. 21 attributes are annotated by [**?**]. 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.

4.1.2. Parameters 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 set to 0. The learning rate is set to 0.01. We used dropout [**?**] for the fully-connected layers with a rate of 0.6.

For tests on APiS and VIPeR, a dimension of 500 is used for the layers fc1 and fc2 as well as for the PCA projected LOMO feature, and the batch size is set to 50. Since the PETA dataset has more data, for tests on PETA, the fc1 layer, fc2 layer and PCA projected LOMO features are set to 1000 dimension and the batch size is set to 100.

The neural network is learned with random initialisation for tests on PETA and APiS. Since for VIPeR we have only 632 training image, the network is pre-trained with triplet loss on CUHK03 dataset[**?**] which contains 13164 images of 1 360 pedestrians. Then, the CNN part is fine-tuned on VIPeR with a lower learning rate (0.0005).

	Accuracy	Recall@FPR=0.1	AUC
LOMO	89.8	72.5	89.8
CNN without part division	90.5	77.3	92.1
CNN with part division	90.8	78.7	92.3
CNN without part division + LOMO	91.5	79.4	91.7
CNN with part division + LOMO	**91.7**	**81.3**	**93.0**

Table 2: Attribute recognition result comparison of different variants of our approach on PETA.

	Accuracy	Recall@FPR=0.2	AUC
non pretrained CNN	81.0	61.5	75.9
pretrained CNN	82.4	65.9	79.3
LOMO	83.1	68.2	**81.0**
pretrained CNN + LOMO	**83.9**	**69.6**	80.9

Table 3: Attribute recognition result comparison of different variants of our approach on VIPeR.

4.1.3. Evaluation measure

The test protocol of PETA [**?**] proposes to use the attribute classification accuracy as evaluation measure. The APiS protocol [**?**] 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) curve as performance measures. As mentioned in [**?**], accuracy is not sufficient to evaluate the classification performance on unbalanced attributes. In our experiments, we thus use all these three measures to evaluate our approach.

4.1.4. Comparison with different variants

We first evaluated the effectiveness of different components: body part division, the feature fusion and person re-identification pre-training. The comparisons among different variants of the method using Peta and VIPeR datasets are respectively shown in Tables **??** and **??**. The body part division improves more than 1% point on the recall score. The fusion of deep features and LOMO features improves about 1% point on accuracy and 2% points on recall score.

Comparing Table **??** and **??**, we can remark that LOMO features are more performant on the VIPeR dataset and the deep features are more performant on the PETA datasets, since the images in VIPeR undergo extreme viewpoint changes, which LOMO is spedifically designed to deal with. The Peta dataset actually comprises several datasets, where various clothing appearances becomes the major aspect of variation. The rich learned feature representation of the CNN is more robust in this case. Finally, the fusion increases the overall recall and accuracy on both datasets. We can conclude that these two kinds of features are complementary and the fusion make the framework more robust.

In Table **??**, we can see that the pre-training on person re-identification data increases the accuracy by 1.4% points , the recall by 4.4% points and the AUC score by 3.4% points. This shows that the capacity to discriminate people learned from person re-identification can assist the attribute learning. In Section **??**, we further show that, inversely, the attributes can assist and improve the person re-identification task.

4.1.5. Comparison with the state-of-the-art methods

The comparison with the state of the art on PETA is shown in Table **??**. In the literature, there are two evaluation settings for the PETA dataset with 35 and 45 attributes respectively. Table **??** 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 on the 27 attributes by 3.4%, 14.3%, 6% points for average accuracy, recall and AUC respectively and by 3.5%, 15%, 6.1% points on the 45 attributes. It also outperforms the DeepMar method by 9% points on accuracy. Moreover, our approach achieves a better score on almost all individual attributes. This superior performance comes from, the fusion of the viewpoint-invariant LOMO features and the rich deep feature representation, on the one hand, and from a better generalisation of our architecture and training compared to the mlcnn approach, on the other hand. We performed a simple test by removing the "dropout" mechanism from the training. Dropout is a

form of regularisation and usually improves the generalisation capacity. On Peta, without dropout, the recall dropped from 79.9% to around 71%, which still above the state of the art.

attribute	Accuracy Rate (%)				Recall@FPR=0.1		AUC	
	MRFr2[?]	DeepMar[?]	mlcnn[?]	ours	mlcnn[?]	ours	mlcnn[?]	ours
personalLess30	**86.8**	85.8	81.1	86.0	63.8	**80.8**	88.5	**93.8**
personalLess45	83.1	81.8	79.9	**84.7**	59.4	**74.9**	84.6	**91.9**
personalLess60	80.1	86.3	92.8	**95.4**	70.2	**83.0**	87.7	**92.8**
personalLarger60	93.8	94.8	97.6	**98.9**	90.7	**94.6**	94.9	**96.8**
carryingBackpack	70.5	82.6	84.3	**85.5**	58.4	**70.2**	85.2	**91.9**
carryingOther	73.0	77.3	80.9	**85.7**	46.9	**65.1**	77.7	**88.4**
lowerBodyCasual	78.2	84.9	90.5	**92.1**	56.2	**76.1**	87.5	**93.1**
upperBodyCasual	78.1	84.4	89.3	**91.2**	62.1	**74.2**	87.2	**92.5**
lowerBodyFormal	79.0	85.2	90.9	**93.3**	72.5	**82.8**	87.8	**92.7**
upperBodyFormal	78.7	85.1	91.1	**93.4**	70.5	**83.4**	87.6	**92.9**
accessoryHat	90.4	91.8	96.1	**97.5**	86.1	**89.9**	92.6	**95**
upperBodyJacket	72.2	79.2	92.3	**94.7**	53.4	**77.4**	81.0	**92.1**
lowerBodyJeans	81.0	85.7	83.1	**87.6**	67.6	**83.2**	87.7	**94.5**
footwearLeatherShoes	87.2	87.3	85.3	**90.2**	72.3	**87.8**	89.8	**95.7**
hairLong	80.1	88.9	88.1	**91.3**	76.5	**88.3**	90.6	**95.6**
personalMale	86.5	**89.9**	84.3	88.9	74.8	**87.0**	91.7	**95.8**
carryingMessengerBag	78.3	82.0	79.6	**84.5**	58.3	**70.7**	82.0	**89.8**
accessoryMuffler	93.7	96.1	97.2	**98.8**	88.4	**93.6**	94.5	**96.2**
accessoryNothing	82.7	85.8	86.1	**89.0**	52.6	**71.5**	86.1	**92.1**
carryingNothing	76.5	83.1	80.1	**84.5**	55.2	**71.8**	83.1	**91.3**
carryingPlasticBags	81.3	87.0	93.5	**96.6**	67.3	**83.6**	86.0	**92.2**
footwearShoes	78.4	80.0	75.8	**80.8**	52.8	**68.3**	81.6	**89.4**
upperBodyShortSleeve	75.8	87.5	88.1	**90.7**	69.2	**86.2**	89.2	**94.5**
footwearSneaker	75.0	78.7	81.8	**85.7**	52.0	**73.0**	83.2	**92.0**
lowerBodyTrousers	82.2	84.3	76.3	**83.4**	56.2	**75.2**	84.2	**92.0**
upperBodyTshirt	71.4	83.0	90.6	**93.3**	63.5	**82.7**	88.7	**92.8**
upperBodyOther	**87.3**	86.1	82.0	86.2	73.2	**80.8**	88.5	**93.5**
27 attributes average	80.8	85.4	86.6	**90.0**	65.6	**79.9**	87.0	**93.0**
35 attr in [**?** **?**] average	75.6	82.6		**91.7**		78.9		**92.0**
45 attr in [**?**] average			87.2	**90.7**	67.3	**82.3**	87.7	**93.8**
53 attributes average				91.7		81.3		93.0

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

attribute	Accuracy	Recall@FPR=0.1			AUC			
	ours	fusion[?]	interact[?]	ours	fusion[?]	interact[?]	DeepMar[?]	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 5: Attribute recognition results on APiS (in %).

The results on the APiS dataset are shown in Table **??**. Our method outperforms the Adaboost approach with fusion features and interaction models by a margin of 6% and 2.3% points respectively in recall at FPR=0.1 and AUC. The Adaboost fusion and interaction methods use simple low-level features like color histograms, LBP features. The improvement of our approach is mainly due to the richer feature presentation of CNN and the horizontal local maximum extraction mechanism of LOMO. For the AUC, DeepMar achieves a slightly better result (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 **??**. 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 to the high intra-class variation and the discriminative power of the pro-

posed part-based CNN architecture.

attribute	Accuracy			Recall@FPR=0.2			AUC
	svm[?]	mlcnn-p[?]	ours	svm[?]	mlcnn-p[?]	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 6: Attribute recognition results on VIPeR (in %).

4.2. Re-identification experiments

4.2.1. Dataset

The CUHK03 dataset [?] includes 13164 images of 1 360 pedestrians and is one of the largest publicly available person re-identification dataset. Each person is taken from two different views. There are two settings labelled with human-annotated bounding boxes and the more challenging detected with automatically generated bounding boxes. In this experiment, we use the latter as this is closer to real-world scenarios. There are 100 identities for test and the

Figure 5: Some example images from CUHK03 datasets.

rest for training and validation, with 20 training/test splits (provided by [**?**]).

We use one camera view as the probe set, and the other as the gallery set. For the gallery, we randomly sample one image for each identity. For the probe set, we use all the images, computing the CMC curve for each of them, and then average over them. This evaluation process is repeated for 20 times and the mean value is reported as the final result.

4.2.2. Training setting

The training is performed in two stages. In the first step, we pre-train the two subnets of the framework. Since the CUHK03 dataset does not have attributes annotations, the attribute network is pre-trained on the PETA dataset as described in section **??**. The identification network is pre-trained with the CUHK03 dataset with 1160 identities in the training set. In the second stage, we remove the output layers of the two subnetworks, and then add a fully-connected layer for the fusion. We train the new fusion layer and fine-tune the rest of the network with lower learning rate.

For the pre-training of the identity network, the learning rate is set to 0.005 and the batch size is set to 100. For the fusion phase, the initial learning rate is set 0.005, and we fine-tune the other part with initial learning rate of 0.0005. The learning rate is then reduced by a factor of 0.7 each 2000 iterations. The weights are initialised from zero-mean Gaussian distribution with a standard deviation of 0.01. We used dropout[**?**] for the fully-connected layers with a rate of 0.7. We generated 50 tuples in each iteration. In each tuple, we randomly

select one reference image and one positive image from the same person but from a different camera view, and 5 negatives images from different persons. In the loss function, the coefficient for the absolute part α is set to 0.02 and the margin is set to 1.

In both training phases, we perform data augmentation by randomly flipping the images and by cropping the center regions with random perturbation. All the inputs are resized to a resolution of 128×48.

4.2.3. Results

Method	rank=1	rank =5	rank =10
FPNN [**?**]	19.9	49.3	64.7
Convnet [**?**]	45.0	75.3	83.4
LOMO+XQDA [**?**]	46.3	78.9	88.6
SS-SVM [**?**]	51.2	80.8	89.6
SI-CI [**?**]	52.2	84.3	92.3
DNS[**?**]	57.3	80.1	88.3
S-lSTM [**?**]	57.3	80.1	88.3
S-CNN SQ [**?**]	61.8	80.9	88.3
CAN[**?**]	63.1	82.9	88.2
ours Identity only	59.7	86.1	93.3
ours fusion Id&Attr	**65.0**	**90.3**	**95.1**

Table 7: Re-identification result on CUHK03 ("detected").

We compare our proposed re-identification approach on the CUHK03(detected) dataset with state-of-art methods Our method achieves the best results on rank 1, rank 5 and rank 10 accuracies. The hard example selection helps the training to converge slightly faster. Fusing the identity with the attributes can improve the results by 5.3, 4.2 points on rank 1 and rank 5. This shows that the identity and attribute information can be complementary for re-identification task and the robustness of the attribute features.

5. Conclusion

In this paper, we have proposed a pedestrian attribute-assisted person re-identification framework. In our approach, the attribute learning is performed

by merging low-level and high-level features learned by a body part-based CNN. By fusing these two kinds of features, the resulting model is robust to spatial variations due to pose or view point changes and incorporates rich feature representations that are able to model the divers appearance of pedestrian attributes. Our attribute recognition model outperforms the state of the art on three public benchmarks. Further, to improve person re-identification, our model uses an improved triplet loss to fuse pedestrian identities and an attribute embedding. We have shown that making use of attributes enhances the re-identification performance. Our final re-identification method achieves the state-of-the-art result on the challenging CUHK03 benchmark.

Aknowledgement

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.