Triplet CNN and Pedestrian Attribute Recognition for Improved Person Re-identification

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

Abstract

In this paper, we propose a pedestrian attribute recognition approach and a CNN-based person re-identification framework enhanced by pedestrian attributes. The knowledge of person attributes can help video surveillance tasks like person re-identification as well as person search, semantic video indexing and retrieval to overcome viewpoint changes with their robustness to the inherent visual appearance variations. Compared to previous approaches, our attribute recognition method using Local Maximal Occurrence (LOMO) features and a Multi-Label Multi-Layer Perceptron (MLMLP) classifier proves to be more robust to different view points and is computationally more efficient. The experiments on three public benchmarks show that the proposed method improves the state-of-the-art on attribute recognition. Furthermore, we integrate our attribute recognition algorithm into a triplet CNN similarity learning framework for person re-identification fusing both learned CNN features and attributes. This fusion leads to an overall improvement, and we achieve state-of-the-art results on person re-identification.

1. Introduction

Recognizing persons is one of the main tasks in video surveillance. As one of the most important cue for human beings to recognize people or objects, visual attributes got a lot of attention recently and have also been used for object recognition [3], action recognition [19], face recognition [10] etc. Pedestrian attributes are defined as semantic mid-level descriptions of people, such as gender, accessories, clothing and so on. The advantage of attributes is that they are more robust to visual changes and that they can be used for “zero-shot” identification. Moreover, other biometric features, like faces, are often not visible or of too low resolution to be useful. The main challenge of visual attribute recognition is the very large intra-class variation. This is due to two reasons. Firstly, attributes, like clothing, can have very diverse appearance. Secondly, images from video surveillance cameras can have drastically different viewing angles. This can lead to very different appearances of the same person and very large spatial shifts of attributes in images. Furthermore, illumination changes, occlusions and low resolution of images make the problem even harder.

Person re-identification consists in matching the same individuals across multiple camera views. The person re-identification task faces similar difficulties as attribute recognition like variations of camera viewpoints, lighting conditions and human pose.

This paper proposes two main contributions:

- An attribute recognition approach based on LOMO features and an MLMLP classifier outperforming the state-of-the-art on three public benchmarks. LOMO features maximize the occurrence in a horizontal stripe forming a representation that is robust to very large horizontal shifts. The MLMLP jointly learns all attributes and thus implicitly their interrelation. Further, our method shows good performance across datasets, i.e. learning on one dataset and testing on another.

- A novel person re-identification framework integrating our attribute recognition network with a CNN to extract strong discriminant low-level features. This combined approach enhances the invariance of the system to viewpoint changes and achieves state-of-the-art results on the CUHK03 dataset.

2. Related work

In the pioneering work of Vaquero et al. [23], mid-level attributes were first used for human recognition. A human parsing technique is employed to segment the regions, and each region is associated with a classifier based on Haar-like features and the dominant colours. The performance
of this approach is rather limited as it depends on the accuracy of the human parsing and requires a frontal image of the person which is not guaranteed in real-world applications. Layne et al. [12] define 15 binary attributes related to clothing, hair style, carried objects and gender. A 2784-dimensional low-level colour and texture feature vector is extracted from each image, and an SVM is trained for each attribute. To exploit the attributes for re-identification, the attribute distance in conjunction with a conventional distance between low-level features such as SDALF [4] is used. The method presented by Zhu et al. [31] extract HSV histograms and MB-LBP and HOG features in the lower body and upper body regions. Adaboost is chosen to perform feature selection and a weighted k-NN for classification. However, for these approaches, the recognition of each attribute is totally independent, i.e. their interrelation is not taken into account. Li et al. [13] proposed two CNN architectures, one for simple and one for multiple attribute recognition. However, the max-pooling layers in CNNs can only guarantee the spatial invariance to some extent. Finally, Zhu et al. [32] proposed to divide the pedestrian images into 15 overlapping parts where each part connects to several CNN pipelines. They further pre-define connections between the parts and the attributes in the fully-connected layers to deal with the shift problem. However, these connections are determined manually, and the model is relatively complex.

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 [14, 30], Local Binary Patterns (LBP) [14, 30] or Gabor features [14]. For example, Gray et al. [6] proposed to use Adaboost to select optimal features among color and texture features. Ma et al. [20] use local descriptors based on color and gradient information and encode them using high-dimensional Fisher vectors. The main metric learning methods include Mahalanobis metrics like KISSME [7], Local Fisher discriminant Analysis (LFDA) [21], Marginal Fisher Analysis (MFA) [27] and Cross-view Quadratic Discriminant Analysis (XQDA) [16].

With the recent success of deep learning for computer vision applications, some convolutional neural network models are proposed for person re-identification. Yi et al. [28] first proposed to apply a Siamese network to person re-identification. To handle geometric problems, DeepReId [15] implements a novel architecture where a patch matching layer models the displacement of body parts. Amed et al. [1] introduced an improved Siamese architecture using the difference of feature maps to measure the similarity. Cheng et al. [2] proposed a variant of the triplet loss function and a CNN network processing parts and the entire body. Varior et al. [24] integrated a gate layer in a Siamese CNN to capture effective subtle patterns in the feature map. And Su et al. [22] proposed a three-stage procedure that pre-trains a CNN with attribute labels of an independent dataset, then fine-tunes the network with ID labels and finally re-trains the network with learned attribute feature embedding on the combined dataset.

3. Proposed Method

3.1. Attribute recognition

The overall procedure of our attribute recognition approach is shown in Fig. 1. Given a cropped pedestrian image, LOMO features are first extracted and projected to a lower-dimensional space, previously learnt by PCA. This feature vector is then classified by an MLMLP, that has been trained off-line on a separate dataset. The output is a vector whose elements represent the scores for each attribute.

3.1.1 LOMO feature

In the LOMO feature proposed by [16], the Retinex algorithm is integrated to produce a colour image that is consistent with human perception. To construct the LOMO features, two scales of Scale Invariant Local Ternary Patterns (SILTP) [17] and an 8×8×8-bin joint HSV histogram are extracted in sliding windows. Following the same procedure, the features are extracted at 3 different 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. The resulting histogram achieves a large invariance to view point changes and, at the same time, captures local region characteristics of a person.

3.1.2 Multi-Label MLP

To classify the extracted features, we propose to use a fully-connected MLP with a hidden layer and Rectified Linear Units (ReLU) [9]. Pedestrian attribute classification is a multi-label problem, i.e. contrary to a standard multi-class classification problem, pedestrian attributes are not mutually exclusive. Further, compared to modern CNN architectures, we have much fewer parameters to learn, which improves its generalisation capacity and reduces the risk of over-fitting and the need of a strong regularisation. For
training the neural network, we use the multi-label version of the sigmoid cross entropy loss:

\[
E = -\frac{1}{N} \sum_{i=1}^{N} \sum_{l=1}^{L} \left[ w_i y_{il} \log(\sigma(x_{il})) + (1 - y_{il}) \log(1 - \sigma(x_{il})) \right]
\]

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

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. In the training set, the positive label appears generally less frequently than the negative one. To handle this imbalance, we added the weight \( w \) to the loss function: \( w = -\log_2(p_i) \), where \( p_i \) is the positive proportion of attribute \( l \) in the dataset.

### 3.2. Attribute-integrated person re-identification

As illustrated in Fig. 2, the proposed person re-identification method uses triplets of examples to train the network with an anchor image \( a \), a positive image \( p \) from the same person as \( a \) and a negative image \( n \) from a different person. The weights of the network for the three input images are shared, and to train the network, the following triplet loss function is minimised:

\[
E_{\text{triplet}} = -\frac{1}{N} \sum_{i=1}^{N} \left[ \max\{\|f(a_i) - f(p_i)\|^2_2, 0\} \right.
\]

\[
\left. - \|f(a_i) - f(n_i)\|^2_2 + m \} \right].
\]

where \( N \) is the number of triplets, \( f \) is the output of the network, and \( m \) is a margin. With the triplet loss function, the network learns a semantic distance metric by "pushing" the negative image pair apart and "pulling" the positive images closer in the feature space. For each input image, there are two branches (see Fig. 3): one for the MLMLP attribute recognition presented in section 3.1, another for a CNN-based low-level feature extraction. In the CNN part, there are three repeated convolution, batch normalization and pooling layers. ReLU activation functions are used. The size of the first convolution is 5×5. The two following are of size 3×3. The kernel size of max-pooling is 2×2, and the number of channels of convolution and pooling layers is 32. Then, similar to [24], we use two layers 1D horizontal convolutions of size 3×1 without zero-padding to reduce the feature maps to a single column. These layers have less parameters and are able to model the displacement in horizontal stripes. Then, the final CNN output vector extracts one feature for each horizontal stripe. In the last convolution layer, the number of channels is increased to 150. This feature is fed to a fully-connected layer to generate an output of 400 dimension. The CNN output and attribute vector are normalized and concatenated. Another fully-connected layer is put on the top of the concatenated vector and learns the optimal fusion of the two representations with output dimension of 400. Dropout [9] is applied to the fully-connected layers to reduce the risk of over-fitting.

### 4. Experiments

In this section, the proposed methods are evaluated on the VIPeR [5] and GRID [18] datasets with the annotation from [11] and the APiS dataset [31] (see Fig. 4). Finally, we test the attribute-integrated triplet CNN for person re-identification on the CUHK03 dataset [15].

#### 4.1. Intra-dataset attribute recognition

The VIPeR dataset [5] contains 632 pedestrian images captured in an outdoor environment, each having 2 images from 2 different view points. GRID [18] contains 1275 pedestrian images captured in an underground station. These two datasets are annotated with 21 attributes by [11]. However, in GRID only 250 pedestrians who have two images from different cameras have attribute annotations. We will only use these images for our attribute recognition experiments. We follow the experiment setting of [32]. All images are scaled to 128x48 pixels, and each dataset is divided into two equal-size disjoint parts for training and testing (images from the same person are not separated). We repeat the process 10 times and report the average result. For Vi-
Table 1. Attribute recognition results on VIPeR

Table 2. Attribute recognition results on GRID

Table 3. Attribute recognition results on APiS
Moreover, the proposed method achieves a good result on GRID and APiS without cross-validating the hyperparameters with this dataset demonstrating a good generalization capacity of our model.

In terms of computational speed, for all images of one trial on VIPeR, the training time is 43.9s and the test time is 2.1s. The LOMO feature extraction for the training and test sets (632 images) takes only 5.6s (< 0.01s/image). For comparison, the CNN approach [32] needs 28.1 minutes for training and 3.6 minutes for test. Thus our system is much more efficient and suitable for real-time applications of video surveillance. And this generates very little extra cost for the proposed re-identification system.

### 4.2. Cross-dataset attribute recognition

In this section, we conduct an experiment in a cross-dataset setting which is more realistic for practical applications. All images in the VIPeR dataset are used for training, and we use the 500 images with attribute annotation in Grid as the test set. We take the same parameter setting in the section 4.1. As the results in Table 4 show, even in the cross-dataset setting, our method can still get a slightly better result than the SVM-based and CNN-based methods trained on the same data set. This demonstrates the excellent generalization capacity of our system.

### 4.3. Person Re-identification

The CUHK03 dataset [15] includes 13,164 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 rest for training and validation, with 20 training/test splits (provided by [15]). Finally, we report the average result over all splits.

Our attribute MLMLP is pre-trained on VIPeR dataset. Then, the training process on CUHK03 is performed in two stages. In the first stage, we train the CNN branch from scratch. In the second stage, we add the attribute branch and the last fully-connected layer. The learning rate is set to 0.01, and we apply a much lower learning rate (0.0005) to the attribute branch for fine-tuning to the CUHK03 dataset. The weights are initialized from zero-mean Gaussian distribution with a standard deviation of 0.01. We randomly generate 50 triplets in each iteration. The margin of triplet loss is set to 1. All the inputs are resized to a resolution 128×48, and we perform data augmentation by randomly flipping the images and by cropping 120×40 regions with random perturbation. For evaluation, we follow the standard protocol and report the one-shot single query Cumulative Match Curve (CMC) on the test set as [15].

The comparison to the state-of-the-art on CUHK03 is shown in Table 5. Our approach achieves the best score at rank 5 and rank 10. And, at rank 1, we are just behind two best methods but still superior to most of the recent state-of-art results. Compared to the baseline, integrating the attributes in the CNN framework could get 1.2% point and 0.7% point improvement on rank 1 and rank 5. This demonstrates the effectiveness of fusing the low-level CNN features and high-level attributes.

<table>
<thead>
<tr>
<th>Method</th>
<th>rank=1</th>
<th>rank=5</th>
<th>rank=10</th>
</tr>
</thead>
<tbody>
<tr>
<td>KISSME [8]</td>
<td>11.7</td>
<td>33.3</td>
<td>48.0</td>
</tr>
<tr>
<td>FPNN [15]</td>
<td>19.9</td>
<td>49.3</td>
<td>64.7</td>
</tr>
<tr>
<td>Convnet [1]</td>
<td>45.0</td>
<td>75.3</td>
<td>83.4</td>
</tr>
<tr>
<td>LOMO+XQDA [16]</td>
<td>46.3</td>
<td>78.9</td>
<td>88.6</td>
</tr>
<tr>
<td>SS-SVM [29]</td>
<td>51.2</td>
<td>80.8</td>
<td>89.6</td>
</tr>
<tr>
<td>Si-CT [26]</td>
<td>52.2</td>
<td>84.3</td>
<td>92.3</td>
</tr>
<tr>
<td>S-ISTM [25]</td>
<td>57.3</td>
<td>80.1</td>
<td>88.3</td>
</tr>
<tr>
<td>S-CNN SQ [24]</td>
<td>61.8</td>
<td>80.9</td>
<td>88.3</td>
</tr>
<tr>
<td>our triplet CNN without attr</td>
<td>53.9</td>
<td>85.4</td>
<td>93.1</td>
</tr>
<tr>
<td>our triplet CNN with attr</td>
<td>55.1</td>
<td>86.1</td>
<td>93.3</td>
</tr>
</tbody>
</table>

Table 5. Re-identification result on CUHK03 (“detected”).
References


