Triplet CNN and pedestrian attribute recognition for improved person re-identification

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

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 2 784-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 convolution 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

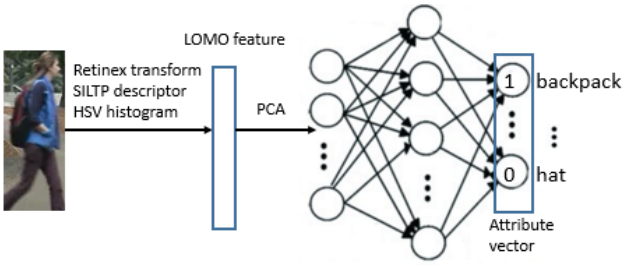

Figure 1. Overview of attribute recognition approach

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 $8\times8\times8$-bin joint HSV histogram are extracted in sliding windows. Following the same procedure, the features are extracted at 3 different scales. For all subwindows on the same image line, only the maximal value of the local occurrence of each pattern among these subwindows 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

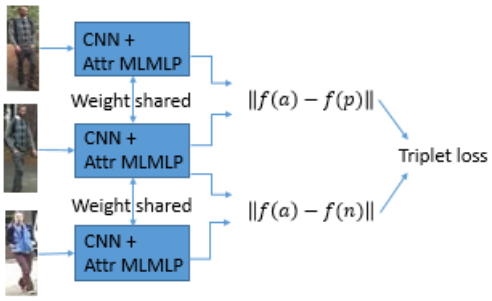

Figure 2. Overview of our person re-identification approach.

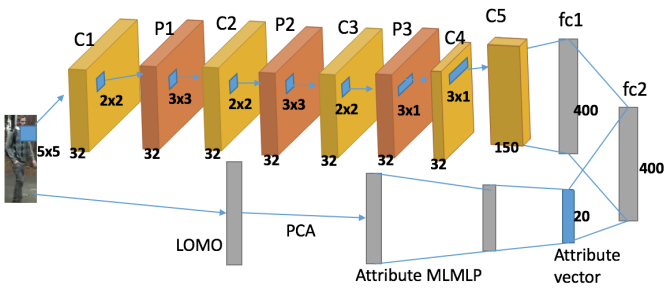

Figure 3. The structure of one branch of the triplet network (blue box in Fig. 2). There are two sub-branches. One is a deep CNN, another is our attribute MLMLP.

training the neural network, we use the multi-label version of the sigmoid cross entropy loss:

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

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

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. 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: $w = -log_2(p_l)$, where p_l 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:

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

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 1 275 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-

Figure 4. Some image examples from pedestrian attribute datasets

PER, one more repetition is performed to determine hyper-parameters like the number of hidden neurons, learning rate and the number of iterations (100, 0.003 and 20 000 in our experiment).

For GRID, the same hyper-parameters are used. The feature vectors are projected into a 500 dimensional sub-space computed by PCA on respectively the VIPeR training images and the GRID training images plus some additional images. For some attributes, there are not enough positive examples like "bald". Thus, we tested 20 attributes in VIPeR and 18 attributes in GRID. We compared to the CNN-based methods in [32] and the SVM-based method in [12] reconstructed by [32]. The accuracy rate with default threshold and the recall with a false positive rate of 0.2 are used as evaluation measures.

The APiS dataset [31] contains 3 661 images, and 11 binary attributes are annotated. We followed the experiment setting of [31]. A 5-fold cross-validation is performed, and the final result is the average of the five tests. We used the same parameter setting as we used in VIPeR dataset, and as performance measure we use the average recall rate at a false positive rate of 0.1.

The results are shown in Tables 1 and 2 on the VIPeR and GRID test sets, Our methods achieves respectively 9% and 6% points improvement in accuracy and 2.4% and 3.4% points on recall compared to the mlcnn-p approach, and even more compared to the SVM approaches. Results on APiS dataset are shown in Table 3, where our approach obtains a 1.2% point improvement on recall compared to the baseline approach of the benchmark which is based on Adaboost and a k-NN classifier. We obtained better results on most of attributes in the three benchmarks. This result demonstrates the robustness against view point changes and the effectiveness of interrelation between attributes of our

attribute	Accuracy Rate (%) average ± std			Recall Rate (%)@FPR=0.2 average ± std		
	SVM	mlcnn-p	Ours	SVM	mlcnn-p	Ours
redshirt	85.5±2.3	91.9±1.0	**94.4**±1.2	88.4±3.9	88.9±4.8	**93.0**±3.2
blueshirt	73.0±5.2	69.1±3.3	**91.6**±1.0	60.8±3.9	70.8±5.1	**72.1**±6.3
lightshirt	83.7±1.0	83.0±1.2	**85.1**±1.3	87.8±1.3	85.3±2.3	**89.6**±1.8
darkshirt	84.2±0.9	82.3±1.4	**84.4**±1.6	87.5±1.2	85.8±2.1	**88.0**±2.2
greenshirt	71.4±5.2	75.9±5.9	**94.7**±1.0	54.3±9.5	69.4±8.0	**71.9**±8.0
nocoat	70.6±1.9	71.3±0.8	**72.9**±1.2	59.3±2.4	57.2±3.2	**61.7**±2.0
notlightdark jeanscolor	70.3±7.3	90.7±2.0	**95.7**±1.1	57.2±7.9	78.6±7.5	**89.1**±3.7
darkbottoms	75.7±1.7	**78.4**±0.7	78.0±1.8	70.2±4.7	**76.2**±1.9	72.9±3.2
lightbottoms	74.7±1.2	76.4±1.2	**77.3**±1.7	69.5±3.0	73.3±2.5	**74.9**±4.1
hassatchel	47.8±4.8	57.8±2.7	**70.2**±1.7	22.0±4.9	31.7±4.3	**39.3**±3.7
barelegs	75.6±3.8	84.1±1.1	**89.5**±1.2	68.7±6.5	85.4±4.5	**92.6**±3.3
shorts	70.4±5.2	81.7±1.3	**90.8**±1.3	59.8±6.5	82.9±4.7	**86.3**±3.3
jeans	76.4±1.3	77.5±0.6	**79.4**±2.0	72.7±3.4	74.7±2.8	**78.8**±4.3
male	66.5±1.1	69.6±2.6	**70.3**±1.2	48.2±3.5	57.2±3.7	**58.6**±2.6
skirt	63.6±8.8	78.1±3.5	**94.4**±0.8	40.7±13.9	60.7±9.9	**71.8**±7.8
patterned	46.9±15.1	57.9±9.2	**90.3**±1.0	26.3±6.0	41.0±9.0	**43.0**±4.8
midhair	64.1±2.3	**76.1**±1.8	73.5±2.1	43.0±3.9	**63.5**±4.2	51.1±4.0
darkhair	63.9±1.8	**73.1**±2.1	67.4±1.2	39.6±2.7	**58.4**±5.8	50.3±3.6
hashandbag carrierbag	45.3±3.8	42.0±6.5	**90.9**±0.8	17.4±3.5	18.5±5.8	**25.9**±6.1
hasbackpack	67.5±1.4	64.9±1.2	**71.3**±1.3	47.9±4.7	49.9±3.7	**53.9**±5.1
average	68.9±1.1	74.1±1.0	**83.1**±0.5	56.1±1.3	65.5±1.5	**68.2**±1.1

Table 1. Attribute recognition results on VIPeR

attribute	Accuracy Rate (%) average ± std			Recall Rate (%)@FPR=0.2 average ± std		
	SVM	mlcnn-p	Ours	SVM	mlcnn-p	Ours
redshirt	74.3±4.9	90.4±2.9	**91.7**±2.7	65.8±10.4	**87.3**±7.1	80.0±4.1
blueshirt	77.8±5.9	84.8±2.8	**92.1**±1.4	70.8±8.9	**85.2**±6.9	72.6±8.1
darkshirt	77.5±2.1	81.2±1.9	**81.6**±2.3	78.1±4.5	**84.4**±5.9	76.7±4.7
darkbottoms	83.8±2.4	83.8±2.6	**84.2**±1.8	86.8±3.7	86.6±4.9	**88.2**±3.6
lightbottoms	83.6±2.3	83.5±2.9	**84.1**±2.3	87.0±4.2	86.8±5.2	**89.1**±2.6
hassatchel	55.4±1.8	55.8±3.6	**65.0**±2.2	29.6±3.6	26.9±4.8	**37.9**±5.7
barelegs	62.0±5.5	76.4±2.4	**82.7**±2.9	40.0±7.6	65.4±5.7	**67.6**±5.3
shorts	62.3±5.5	67.4±5.0	**86.0**±2.2	39.5±10.5	22.0±5.5	**61.1**±8.6
jeans	60.6±3.1	62.4±1.8	**66.4**±2.1	40.7±5.4	42.2±6.9	**50.0**±5.3
male	63.2±2.9	68.4±1.8	**70.2**±2.7	42.8±8.2	52.8±4.9	**57.1**±6.1
skirt	27.0±31.7	73.8±4.9	**88.4**±2.2	17.3±5.7	44.4±13.6	**60.5**±10.3
patterned	58.5±13.7	74.3±3.3	**86.1**±1.9	38.3±13.7	44.7±13.3	**45.1**±7.1
midhair	61.1±2.8	72.4±3.4	**73.6**±3.4	38.4±8.5	**60.9**±7.8	49.9±8.5
darkhair	59.6±5.0	**71.8**±3.6	71.7±2.7	37.6±9.3	**58.3**±6.2	49.8±7.7
hashandbag carrierbag	54.6±8.8	61.8±2.8	**70.1**±3.8	30.1±5.1	34.7±8.5	**45.0**±9.5
hasbackpack	61.8±2.4	63.1±3.4	**73.0**±3.4	43.3±3.4	33.7±6.2	**39.9**±6.9
16attributes average	63.9±2.3	73.2±0.7	**79.3**±0.4	49.1±1.8	57.3±0.9	**60.7**±2.2
lightshirt			78.5±3.0			77.5±5.9
nocoat			82.2±4.2			51.9±20.1
18attributes average			**79.2**±0.5			**61.1**±1.8

Table 2. Attribute recognition results on GRID

attribute	Recall Rate (%)@FPR=0.1		Accuracy Rate (%)
	Baseline	Ours	Ours
T-shirt	**55.22**	50.60	73.56
backpack	**56.16**	51.74	89.27
gender	**58.30**	50.08	74.11
hand carrying	52.14	**53.68**	84.46
longhair	55.15	**56.10**	86.64
longjeans	**89.85**	89.15	67.71
longpants	76.68	**84.41**	89.89
M-s Pants	78.65	**85.71**	89.76
shirt	54.62	**56.37**	83.17
s-s bag	38.45	**41.49**	78.04
skirt	68.23	**73.52**	93.12
average	61.75	**62.99**	82.70

Table 3. Attribute recognition results on APiS

	Training set	Test set	recall	accuracy
SVM	GRID	GRID	49.1	63.9
mlcnn-p	GRID	GRID	57.3	73.2
ours	VIPeR	GRID	**57.8**	**73.9**

Table 4. Cross-dataset attribute recognition results.

MLMLP-based method.

Moreover, the proposed method achieves a good result on GRID and APiS without cross-validating the hyper-parameters with this dataset demonstrating a good generalization capacity of our model.

In terms of computational speed, for all images of one trial on the 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 (0,000 5) 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].

Method	rank=1	rank =5	rank =10
KISSME [8]	11.7	33.3	48.0
FPNN [15]	19.9	49.3	64.7
Convnet [1]	45.0	75.3	83.4
LOMO+XQDA [16]	46.3	78.9	88.6
SS-SVM [29]	51.2	80.8	89.6
SI-CI [26]	52.2	84.3	92.3
S-lSTM [25]	57.3	80.1	88.3
S-CNN SQ [24]	**61.8**	80.9	88.3
our triplet CNN without attr	53.9	85.4	93.1
our triplet CNN with attr	55.1	**86.1**	**93.3**

Table 5. Re-identification result on CUHK03 ("detected").

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.

5. Conclusion

In this paper, a pedestrian attribute classification approach based on LOMO features and a Multi-Label MLP has been proposed. This approach has the properties of both being robust to large view point variations as well as being computationally efficient. We performed experiments on three public datasets and outperformed the state-of-art methods. We further proposed a framework for person re-identification integrating our attribute recognition method with a triplet CNN similarity metric learning architecture. We obtained results that are equivalent or superior to most state-of-the-art re-identification methods, and show that the high-level attribute information can help improving person re-identification with low-level features.

Acknowledgement

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.

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *Computer Vision and Pattern Recognition*, pages 3908–3916, 2015.

[2] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Computer Vision and Pattern Recognition*, pages 1335–1344, 2016.

[3] K. Duan, D. Parikh, D. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *Computer Vision and Pattern Recognition*, 2012.

[4] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Computer Vision and Pattern Recognition*, pages 2360–2367, 2010.

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

[6] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European Conference on Computer Vision*, pages 262–275, 2008.

[7] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *Computer Vision and Pattern Recognition*, pages 2288–2295, 2012.

[8] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *Computer Vision and Pattern Recognition*, pages 2288–2295, 2012.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[10] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *International Conference on Computer Vision*, 2009.

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

[12] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary. Person re-identification by attributes. In *British Machine Vision Conference (BMVC)*, 2012.

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

[14] W. Li and X. Wang. Locally aligned feature transforms across views. In *Computer Vision and Pattern Recognition*, pages 3594–3601, 2013.

[15] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid:deep filter pairing neural network for person re-identification. In *Computer Vision and Pattern Recognition*, pages 152–159, 2014.

[16] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Computer Vision and Pattern Recognition*, pages 2197–2206, 2015.

[17] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikäinen, and S. Z. Li. Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In *Computer Vision and Pattern Recognition*, pages 1301–1306, 2010.

[18] C. Liu, S. Gong, C. C. Loy, and X. Lin. Person re-identification: What features are important? In *European Conference on Computer Vision*, pages 391–401, 2012.

[19] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *Computer Vision and Pattern Recognition*, pages 3337–3344, 2011.

[20] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *European Conference on Computer Vision*, pages 413–422, 2012.

[21] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *Computer Vision and Pattern Recognition*, pages 3318–3325, 2013.

[22] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian. Deep attributes driven multi-camera person re-identification. In *European Conference on Computer Vision*, pages 475–491, 2016.

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

[24] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *European Conference on Computer Vision*, pages 791–808, 2016.

[25] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *European Conference on Computer Vision*, pages 135–153, 2016.

[26] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *Computer Vision and Pattern Recognition*, pages 1288–1296, 2016.

[27] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *European Conference on Computer Vision*, 2014.

[28] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Deep metric learning for person re-identification. In *International Conference on Pattern Recognition*, pages 34–39, 2014.

[29] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan. Sample-specific svm learning for person re-identification. In *Computer Vision and Pattern Recognition*, pages 1278–1287, 2016.

[30] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *International Conference on Computer Vision*, pages 2528–2535, 2013.

[31] J. Zhu, S. Liao, Z. Lei, D. Yi, and S. Li. Pedestrian attribute classification in surveillance: Database and evaluation. In *International Conference on Computer Vision Workshops*, pages 331–338, 2013.

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