POSE GUIDED DEEP MODEL FOR PEDESTRIAN ATTRIBUTE RECOGNITION IN SURVEILLANCE SCENARIOS

Dangwei Li[1,2]*, Xiaotang Chen*[1,2]*, Zhang Zhang*[1,2]*, and Kaiqi Huang*[1,2,3]

[1]CRIPAC & NLPR, CASIA [2]University of Chinese Academy of Sciences
[3]CAS Center for Excellence in Brain Science and Intelligence Technology
{dangwei.li, xtchen, zzhang, kaiqi.huang}@nlpr.ia.ac.cn

ABSTRACT

Recognizing pedestrian attributes, such as gender, backpack, and cloth types, has obtained increasing attention recently due to its great potential in intelligent video surveillance. Existing methods usually solve it with end-to-end multi-label deep neural networks, while the structure knowledge of pedestrian body has been little utilized. Considering that attributes have strong spatial correlations with human structures, *e.g*. glasses are around the head, in this paper, we introduce pedestrian body structure into this task and propose a Pose Guided Deep Model (PGDM) to improve attribute recognition. The PGDM consists of three main components: 1) coarse pose estimation which distillates the pose knowledge from a pre-trained pose estimation model, 2) body parts localization which adaptively locates informative image regions with only image-level supervision, 3) multiple features fusion which combines the part-based features for attribute recognition. In the inference stage, we fuse the part-based PGDM results with global body based results for final attribute prediction and the performance can be consistently improved. Compared with state-of-the-art models, the performances on three large-scale pedestrian attribute datasets, *i.e*., PETA, RAP, and PA-100K, demonstrate the effectiveness of the proposed method.

Index Terms— Pedestrian attribute recognition, intelligent video surveillance, person retrieval, pose estimation

1. INTRODUCTION

Recognition of pedestrian attributes, *e.g*. gender, backpack, cloth types shown in Fig. 1, has obtained increasing attention recently due to its great potential in real applications, such as person re-identification and attribute-based person retrieval in intelligent video surveillance. For example, describable person attributes can play a critical role for the search of the two suspects in Boston marathon bombing event [1]. Despite of years of efforts, there still are many challenges, such as pose variation, illumination variation, camera viewing angle, and the low-quality image due to far distance cameras.

Existing methods for pedestrian attribute recognition typically consist of two stages, *i.e*. feature representation and attribute classifiers design. In the past, researchers usually adopt hand-crafted features, *e.g*. Ensemble of Local Features (ELF) [2], then learn classifiers for each attribute separately [3]. Recently, deep learning, especially Deep Convolutional Neural Networks (DCNN), has made great progress on general object recognition [4, 5]. Inspired by this, researchers introduce the DCNN into pedestrian attribute recognition [6–8]. They treat it as an end-to-end multi-label classification task and achieve considerable improvements compared with previous two-stage based methods. Thus, we also utilize DCNN to solve pedestrian attribute recognition task.

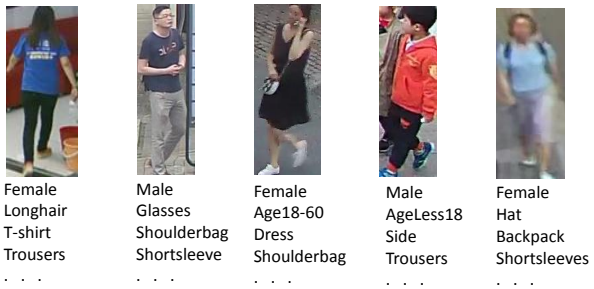

Fig. 1. Examples of pedestrian attributes in real scenarios.

Typically, pedestrian attributes own strong spatial relationships with human body parts. For example, hair types are around the head. Shoe types can be determined by the region of foot. Existing methods typically explore simple rigid structure by slicing the human image into multiple rigid strides or blocks [8, 9]. However, those rigid partitions cannot well depict the pedestrian pose variation, and also partially damage the attribute structure, *e.g*. breaking the Tshirt into two parts. Differently, in this paper, we explore the human body structure, *i.e*. pedestrian pose, for pedestrian attribute recognition.

To utilize the pedestrian pose for attribute recognition, we should solve two basic problems. The first one is human pose estimation. To our knowledge, there are no pose annotations in existing pedestrian attribute datasets. Besides, re-annotating human poses on existing pedestrian attribute datasets is also a hard and costly project. The second one is how to apply the pose knowledge to attribute recognition. Common pose estimation can only produce human key points, and attributes are typically corresponded with regions. How

978-1-5386-1737-3/18/$31.00

to build the relationships between points and regions is still a problem. In this paper, we propose a Pose Guided Deep Model (PGDM) to jointly solve both the problems. For the first problem, considering that there are already some good pose estimation models [10], instead of re-annotation, the PGDM transfers those pose knowledge from existing pose estimation models to pedestrian attribute datasets. For the second problem, based on the prior key points, the PGDM discovers a suitable part region around each key point, and then ensembles all key point related regions for pedestrian attribute recognition. To be noticed, the pose knowledge distillation and region localization in PGDM are optimized jointly.

In the inference stage, the PGDM first estimates the human key points and generates the part regions simultaneously, then ensembles these region-based feature representations for pose-guided pedestrian attribute recognition. In addition to the region-based predictions produced by PGDM, we also obtain the global body-based predictions and fuse these two results at score level as the final prediction. In summary, the contributions of this paper include:

- To our knowledge, this is the first attempt to explore the deformable pedestrian body structure knowledge, *e.g.* pose information, for pedestrian attribute recognition.
- A pose guided deep model is proposed, which could not only transfer the pose knowledge from existing pose estimation model to surveillance scenarios, but also adaptively locate informative regions for the high-level attribute recognition task.
- The proposed method has obtained competitive results in three large-scale pedestrian attribute datasets.

2. RELATED WORK

Early works typically adopt classical hand-crafted features and train multiple binary classifiers for each attribute independently. Layne et al. [3] first propose to train Support Vector Machines (SVM) classifiers to recognize human attributes and utilize the recognition results to assist person re-identification based on ELF features. Zhu et al. [11] utilize Gentle AdaBoost algorithms with multiple features, *e.g.* color feature, HOG feature, to jointly make feature selection and attribute classification. Deng et al. [12] introduce the intersection kernel SVM for attribute recognition and use Markov Random Field (MRF) as post processing.

Recently many researchers utilize deep learning to solve pedestrian attribute recognition due to its great power in feature learning, which may better handle the complex variations in surveillance scenes. Li et al. [6] treat pedestrian attribute recognition as a multi-label classification problem and propose an improved entropy loss to handle unbalance label distribution. Sudowe et al. [13] propose the Attribute Convolutional Net (ACN) to jointly learn different attributes through a jointly-trained holistic CNN model. Zhou et al. [14] utilize

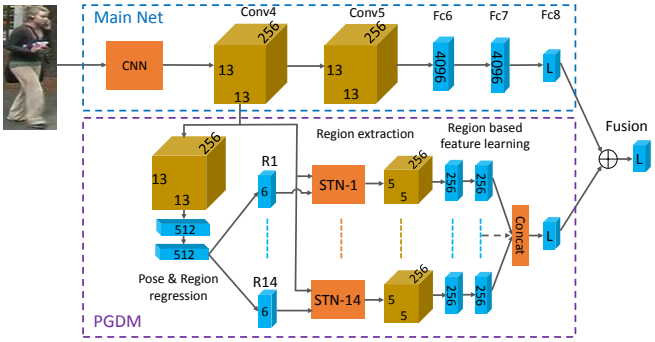

Fig. 2. Overall framework. (Top) The Main Net learns one independent classifier for each attribute and the output of 'Conv4' layer are used for PGDM. (Bottom) The PGDM first regresses 6 transformer parameters (including 2 pose key point position) for each region, then learn representation for each region produced by spatial transformer network, and lastly ensemble 14 regions' representation for region-based attribute recognition. Max pooling layers after the last convolutional layers in the Main Net and regression network of PGDM are omitted. Best viewed in color.

spatial pyramid pooling to jointly handle attribute recognition and localization.

In addition to these straightforward methods, some researchers also explore human prior structure to assist pedestrian attribute recognition. Zhu et al. [9] first divide human body into 15 rigid parts and train a CNN model for each part, then use prior fully connected layer to fuse multiple parts for attribute recognition. Sarfraz et al. [15] utilize human view angle to assist the recognition of pedestrian attributes. Yao et al. [16] develop an adaptive region localization method for attribute recognition. Wang et al. [8] explore pedestrian rigid strides to capture spatial information for attribute recognition. Liu et al. [7] utilize multiple attention models to simultaneously learn multi-level feature representation for attribute recognition. Different from those methods, in this paper, we explore the pedestrian deformable body structure knowledge, *i.e.* human pose, to improve pedestrian attribute recognition.

3. PROPOSED METHOD

In this paper, we propose a deep convolutional neural network for pedestrian attribute recognition, which adaptively learns informative regions through exploring pedestrian body structure knowledge. The overall framework is shown in Fig. 2, which consists of two components, Main Net and Pose Guided Deep Model. The details about these two components are described as follow.

3.1. Main Net

The Main Net follows the structure of CaffeNet, which has little modification of AlexNet [4], and the basic structure is shown on Fig. 2. For the Main Net, we treat the pedestrian attribute recognition as a multi-label classification problem, and

use improved cross entropy loss [6] as our objective function, which could partially handle the unbalanced label distribution in different attribute categories. Considering there are L attributes and N images, the optimization objective can be formalized as follows:

$$L_m = -\frac{1}{NL}\sum_{i=1}^{N}\sum_{l=1}^{L} w_l(y_{il}log(\hat{y}_{il}) + (1-y_{il})log(1-\hat{y}_{il})) \tag{1}$$

$$w_l = \begin{cases} \exp((1-p_l)/\sigma^2) & y_{il}=1 \\ \exp(p_l/\sigma^2) & y_{il}=0 \end{cases} \tag{2}$$

where y_{il} is the ground truth of l-th attribute of i-th sample, p_{il} is the corresponding prediction probability, w_l is the loss weight for l-th attribute, p_l is the positive ratio of l-th attribute in the training set, and σ is a temperature coefficient which is set as 1 in this paper.

3.2. Pose Guided Deep Model

The Pose Guided Deep Model (PGDM) aims to explore the deformable body structure knowledge, *i.e*. human pose, to assist pedestrian attribute recognition. It consists of three main components, including coarse pose estimation, adaptively region localization and region-based feature ensemble for attribute recognition.

Coarse Pose Estimation: To our knowledge, there are no human pose annotations in existing pedestrian attribute datasets. Re-annotating pose information on existing attribute datasets is another challenging problem, which is costly and hard due to the low image quality. Due to the deep learning technologies and large-scale person pose datasets, *e.g*. MPII [17] and Leeds Sports Pose (LSP) [18], there are already some good deep pose estimation models, *e.g*. Convolutional Pose Machines (CPM) [10] which has shown well generalization ability and has also been for pose alignment in related re-identification task [19, 20].

In this paper, instead of re-annotation the human pose, we transfer the pose knowledge from people in generic scenarios to pedestrians in surveillance scenarios. Different from the work [19] which uses an extern pose estimation model, we embed the pose estimation model into the pedestrian attribute model for fast inference in test stage. Specifically, first we adopt CPM model [10], which is trained on MPII and LSP with six stages, to generate 14 prior human pose key points as well as confidence scores. The human key points consist head, neck, right shoulder, right elbow, right wrist, left shoulder, left elbow, left wrist, right hip, right knee, right ankle, left hip, left knee, left ankle. Second, taking the generated pose key points as coarse ground truth pose, we can train a regression network to regress pedestrian pose. To be noticed, the pose regression network has shared parameters with region regression network. In this paper, we adopt the smooth-L1 loss [21] with pose prior probability as the objective function for pose regression, which is described as follows:

$$L_r = \frac{1}{2NK}\sum_{i=1}^{N}\sum_{k=1}^{K} S_{i,k}(smooth_{L_1}(\hat{X}_{i,k}-X_{i,k}) + smooth_{L_1}(\hat{Y}_{i,k}-Y_{i,k})) \tag{3}$$

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2 & if|x|<1 \\ |x|-0.5 & otherwise \end{cases} \tag{4}$$

where the $(X_{i,k}, Y_{i,k})$ are the normalized position (ranging in [-1, 1]) of *k-th* key point of *i-th* sample. $S_{i,k}$ is the confidence score of *k-th* key point of *i-th* sample generated from original CPM model. Here, we use the confidence scores as the weight in the pose regression model, which partially makes the model to be robust to the noisy of extracted pseudo ground truth.

Adaptively Region Localization: Pose information is commonly represented as a set of key points, while pedestrian attributes typically correspond to different regions. To transform the key points into informative regions, we propose to regress a bounding box for each key point. Here, we use Spatial Transformer Networks (STN) [22] for region extraction. The STN has shown superior ability in modeling the variance of scale and pose in different tasks, which is suitable for adaptively region localization. There are two components in STN, *i.e*. spatial localization network for affine parameters regression and grid generator to sample the input image using an image interpolation kernel. In our work, spatial localization network is shown as "region regression" in Fig. 2, which outputs six parameters for each key point from R1 to R14. With the affine parameters θ, the grid generator can produce the position mapping as follows:

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \tag{5}$$

where the (x_i^t, y_i^t) are the target coordinates of regular grid in the output feature map and the (x_i^s, y_i^s) are the source coordinates in the input feature map that define the sample points. Based on the position mapping output of grid generator, the STN interpolates interesting region based on bilinear kernel. For more details about STN, please see [22]. To be noticed, for each key point, the normalized position $(X_{i,k}, Y_{i,k})$ is exactly the same to the normalized bias $(\theta_{13}, \theta_{23})$ of region k of *i-th* image, and we use two symbols for easy to present.

Region-based Features Ensemble: To integrate different regions for high-level tasks, for each key point related region, an independent neural network is used for feature learning. Specially, we use two fully connected layer with 256-dimension embedding to learn features of each region, thus the final region-based representation is 3,584-dimension (256×14). Based on the ensemble feature representation from 14 regions, we use a fully connected layer with L outputs as the classifiers to classify all the L attributes. The same objective function as the Main Net is used to optimize the region-based attribute recognition, which is denoted as L_p.

3.3. Optimization and Inference

The final optimization objective is denoted as follows:

$$L = \lambda_1 L_m + \lambda_2 L_p + \lambda_3 L_r \quad (6)$$

where the $\lambda_1, \lambda_2, \lambda_3$ represent the loss weights of Main Net, PGDM and pose regression, respectively. The network is trained in three stages. First, we initialize the main network with pretrained weights on ILSVRC 2012 datasets and train the main network with $\lambda_1 = 1$ and $\lambda_2 = \lambda_3 = 0$. Second, we train the PGDM with $\lambda_1 = 0$, $\lambda_2 = 1$ and $\lambda_3 = 10$, and the main network is freezed. Third, we train the overall network with $\lambda_1 = 10$, $\lambda_2 = 1$ and $\lambda_3 = 1$. To be noticed, the learning rate of main network and region regression network are $1/10$ of the region feature learning. In the inference stage, we fuse the scores of Main Net and PGDM as the final results.

Our model is implemented based on Caffe [23]. As the dataset can be quite large, we use a stochastic approximation of the objective function. The training data is randomly divided into mini-batches with a batch size of 64. We start with a initial learning rate of $\eta = 1 \times 10^{-3}$ and gradually decrease it after each 2×10^4 iterations. Specially, for the third stage, the base learning rate is $\eta = 1 \times 10^{-4}$. We use a momentum of $\mu = 0.9$ and weight decay $\gamma = 5 \times 10^{-3}$. We use the model at 5×10^4 iterations for testing. For image preprocessing, the channel-based mean subtraction and random horizonal mirror are used to prevent overfitting.

4. EXPERIMENTS

4.1. Datasets

We evaluate the proposed method on current three large-scale pedestrian attribute datasets, including PETA [12], RAP [24] and PA-100K [7].

PETA contains 19,000 pedestrian images which are collected from existing person re-identification datasets. It is labeled with 61 binary attributes and 4 multi-class attributes. As previous work [12], 35 attributes are selected for evaluation due to the unbalanced attribute distribution. We adopt the five times random partitions provided by the work [6] for a fair comparison, including 9,500 samples for training, 1,900 samples for evaluation, and 7,600 samples for testing.

RAP contains 41,585 images which are collected from an indoor camera network. It contains 72 fine-grained attributes as well as view angles, occlusion patterns and coarse pedestrian parts. As the work [24], we adopt the five random data partitions for a fair comparison. There are 33,268 images for training and 8,317 images for testing in each partition.

PA-100K contains 100,000 images from 598 outdoor scenes. It has 26 common attributes, including global attributes, such as gender, and object-level attributes, such as backpack. Here, we use the public dataset partition provided by the work [7], which include 80,000 images for training, 10,000 for validation and 10,000 for testing, for evaluation.

Evaluation: We adopt two kinds of metrics for a fair comparison. The first one is the label-based metric, *i.e*. mean accuracy (mA). It first computes each attribute's accuracy, which is the mean recognition rate of positive and negative samples, then makes an average over all the attributes. The second is the instance-based metric, *e.g*. accuracy, recall rate, precision and F1 score. The instance-based metric may be more suitable for person retrieval in real applications, which evaluates the results of multiple concurrent attributes on each sample. The metrics are the same as the work [24].

4.2. Experimental results

We compare the proposed method with current state-of-the-art methods on PETA, RAP and PA-100K. In summary, these methods are grouped into two categories. The first one is two-stage based methods, *e.g*. extracting hand-crafted features (such as ELF) or deep features (FC6 from CaffeNet [23]), and then learning SVM classifier for each attribute. The second one is end-to-end deep learning based methods, which learn feature representation and classifiers jointly. It includes ACN [13], Deep Multi-attribute Recognition (DeepMAR [6]), Flexible Spatial Pyramid Pooling (FSPP [14]), Contextual CNN-RNN (CTX [25]), Semantic Regularisation (SR [26]), Joint Recurrent Learning of context and correlation (JRL [8]), and Hydra-Plus (HP-Net [7]).

Methods	label-based	instance-based			
	mA	Accuracy	Precision	Recall	F1
MRFr2 [12]	75.6	-	-	-	-
ELF+SVM [24]	75.21	43.68	49.45	74.24	59.36
FC7+SVM [24]	72.28	31.72	35.75	71.78	47.73
FC6+SVM [24]	73.32	33.37	37.57	73.23	49.66
ACN [13]	81.15	73.66	84.06	81.26	82.64
DeepMAR [6]	82.89	75.07	83.68	83.14	83.41
FSPP [14]	81.67	75.72	84.84	83.10	83.96
CTX [25]	80.13	-	79.68	80.24	79.68
SR [26]	82.83	-	82.54	82.76	82.65
JRL [8]	**85.67**	-	**86.03**	**85.34**	**85.42**
HP-Net [7]	81.77	76.13	84.92	83.24	84.07
Main Net	82.78	**76.87**	85.30	84.22	84.76
PGDM	82.27	76.57	85.29	84.09	84.69
Fusion	**82.97**	**78.08**	**86.86**	**84.68**	**85.76**

Table 1. Experimental results on PETA. In each column, the 1^{st} and 2^{nd} best results (%) are indicated in **bold**.

PETA: As shown in Table 1, compared with GoogLeNet based methods, *e.g*. FSPP, HP-Net, and VGG based methods, *e.g*. CTX, SR, our fusion method which is based on shallower network CaffeNet, has achieved better results on *instance-based* evaluation. According to the lable-based metric, the JRL has achieved the highest mA score, where 10 variant models based on CaffeNet are combined for a superior performance. Our fusion model achieves the second best performance with the mA score. However, the proposed method still has 0.34% improvement on *F1*. Compared with the single model result of JRL (82.13% [8]), we could still obtain 0.86% improvements than JRL on *mA*.

RAP: As shown in Table 2, we find that the proposed method doesn't achieve better results than existing state-of-

the-art methods, *e.g*. FSPP and HP-Net, and only obtains comparable results on *Accuracy* and *Precision*. We think the main reason may be the large number of occlusion images in RAP (32.3%) as the dataset are collected in indoor scenes. Due to the occlusions of person body, some wrong estimations of key points may decrease the robustness of region-based features. Moreover, the representation ability of CaffeNet is not as well as GoogLeNet to handle the complex variation. So the FSPP and HP-Net can achieve better results on RAP.

Methods	label-based	instance-based			
	mA	Accuracy	Precision	Recall	F1
ELF+SVM [24]	69.94	29.29	32.84	71.18	44.95
FC7+SVM [24]	72.28	31.72	35.75	71.78	47.73
FC6+SVM [24]	73.32	33.37	37.57	73.23	49.66
ACN [13]	69.66	62.61	**80.12**	72.26	75.98
DeepMAR [6]	73.79	62.02	74.92	76.21	75.56
FSPP [14]	**79.64**	60.25	69.10	**80.16**	74.21
CTX [25]	70.13	-	71.03	71.20	70.23
SR [26]	74.21	-	75.11	76.52	75.83
JRL [8]	**77.81**	-	78.11	**78.98**	**78.58**
HP-Net [7]	76.12	**65.39**	77.33	78.79	**78.05**
Main Net	73.77	62.59	76.76	74.77	75.75
PGDM	74.24	62.22	75.75	75.56	75.66
Fusion	74.31	**64.57**	**78.86**	75.90	77.35

Table 2. Experimental results on RAP. In each column, the 1^{st} and 2^{nd} best results (%) are indicated in **bold**.

PA-100K: As shown in Table 3, compared with the HP-Net which is based on GoogLeNet with Batch Normalization [27], the proposed CaffeNet based method has obtained better performance on both *label-based* and *instance-based* metrics. Note that, in PA-100K, the multiple image samples belonging to the same person will not appear in both train and test sets at the same time, which is more close to real scenarios and much challenging.

Methods	label-based	instance-based			
	mA	Accuracy	Precision	Recall	F1
DeepMAR [6]	72.70	70.39	82.24	80.42	81.32
M-Net [7]	72.30	70.44	81.70	81.05	81.38
HP-Net [7]	74.21	**72.19**	82.97	82.09	**82.53**
Main Net	73.92	71.26	**82.98**	80.93	81.94
PGDM	**75.01**	70.67	80.91	**82.34**	81.64
Fusion	**74.95**	**73.08**	**84.36**	**82.24**	**83.29**

Table 3. Experimental results on PA-100K. In each column, the 1^{st} and 2^{nd} best results (%) are indicated in **bold**.

Totally, as we can see from the results on three datasets, the proposed PGDM can obtain complementary representations to the Main Net. The fusion of Main Net and PGDM can substantively improve the final prediction results.

4.3. Ablation Study

To explore what's the difference learned between the Main Net and PGDM, we make a statistic on the attribute categories in terms of recognition results. We select partial attributes' recognition results where PGDM has obtained better results than Main Net, and the results are shown in Table 4. It is obvious that after utilizing the human semantic structure knowledge, the PGDM can recognize some attributes which correspond with human key points better. We also visualize the discovered regions corresponding to each key point in Fig. 3, where the configurations of the regions can be changed adaptively according to the pose variations of pedestrian.

Attribute	Galsses	HandBag	HoldObject	UpperLogo	LowerStripe
Main Net	72.67	67.08	49.90	74.92	71.99
PGDM	74.48	68.37	52.26	77.69	79.82

Table 4. Partial attribute recognition result on PA-100K.

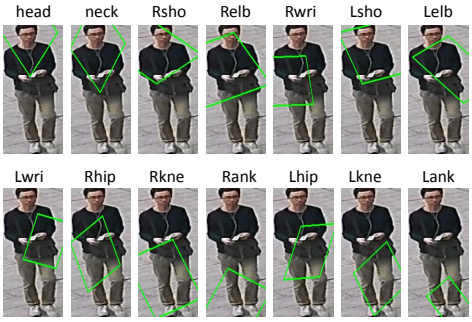

Fig. 3. Visualization of the learned region for each key point. Full name of pose points are described in Section 3.2.

4.4. Attribute based Person Retrieval

As an core application of pedestrian attribute recognition, pedestrian retrieval with multi-attribute query has been tested in this work. Based on the proposed model, we make some visualization about the attribute-driven pedestrian retrieval on PA-100K test set. We use the product of multiple query attributes' predicted probabilities as the final similarity between the image and the query attributes. The experimental results are shown in Fig. 4. We find that even with complex multiple query with 4 attribute categories, the retrieval precisions of some query conditions are still very high in top ranks. Moreover, some images from the same person are discovered in the top ranks, which shows the potential of attribute recognition to assist person re-identification task.

5. CONCLUSION

In this paper, we have introduced the pedestrian structure knowledge into pedestrian attribute recognition task and proposed a pose guided deep model to improve attribute recognition. Experimental results have shown that the proposed PGDM can produce complementary results to global body-based Main Net, and the fusion of Main Net and PGDM can produce better results. In the future, we will explore more efficient strategies to utilize the human semantic structure knowledge to assist pedestrian attribute recognition.

6. ACKNOWLEDGEMENT

This work is jointly supported by the National Key Research and Development Program of China (2016YFB1001005), the National Natural Science Foundation of China (Grant No.

Fig. 4. Visualization of multi-attribute based person retrieval. "nGT" is the number of true positives among gallery set (10,000 images). "AP" is the average precision of nGT true positive samples. The red and green boxes represent true and negative positive samples, respectively. Best viewed in color.

61473290, Grant No. 61673375), the Projects of Chinese Academy of Science (Grant No. QYZDB-SSW-JSC006, Grant No.173211KYSB20160008), and Huawei Technologies Co., Ltd (Contract No.:YBN2017030069).

7. REFERENCES

[1] Rogerio Feris, Russel Bobbitt, Lisa Brown, and Sharath Pankanti, "Attribute-based people search: Lessons learnt from a practical surveillance system," in *ICMR*, 2014, p. 153.

[2] Douglas Gray and Hai Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," *ECCV*, pp. 262–275, 2008.

[3] Ryan Layne, Timothy M Hospedales, Shaogang Gong, and Q Mary, "Person re-identification by attributes," in *BMVC*, 2012.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[5] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015, pp. 1–9.

[6] Dangwei Li, Xiaotang Chen, and Kaiqi Huang, "Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios," in *ACPR*, 2015, pp. 111–115.

[7] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang, "Hydraplus-net: Attentive deep features for pedestrian analysis," in *ICCV*, 2017.

[8] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li, "Attribute recognition by joint recurrent learning of context and correlation," in *ICCV*, 2017.

[9] Jianqing Zhu, Shengcai Liao, Dong Yi, Zhen Lei, and Stan Z Li, "Multi-label cnn based pedestrian attribute learning for soft biometrics," in *ICB*, 2015, pp. 535–540.

[10] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh, "Convolutional pose machines," in *CVPR*, 2016, pp. 4724–4732.

[11] Jianqing Zhu, Shengcai Liao, Zhen Lei, Dong Yi, and Stan Li, "Pedestrian attribute classification in surveillance: Database and evaluation," in *ICCV Workshops*, 2013, pp. 331–338.

[12] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang, "Pedestrian attribute recognition at far distance," in *ACM Multimedia*, 2014, pp. 789–792.

[13] Patrick Sudowe, Hannah Spitzer, and Bastian Leibe, "Person attribute recognition with a jointly-trained holistic cnn model," in *CVPR Workshops*, 2015, pp. 87–95.

[14] Yang Zhou, Kai Yu, Biao Leng, Zhang Zhang, Dangwei Li, Kaiqi Huang, Bailan Feng, and Chunfeng Yao, "Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization," in *BMVC*, 2017.

[15] M Saquib Sarfraz, Arne Schumann, Yan Wang, and Rainer Stiefelhagen, "Deep view-sensitive pedestrian attribute inference in an end-to-end model," in *BMVC*, 2017.

[16] Chunfeng Yao, Bailan Feng, Defeng Li, and Jian Li, "Hierarchical pedestrian attribute recognition based on adaptive region localization," in *ICME Workshops*, 2017, pp. 471–476.

[17] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *CVPR*, 2014, pp. 3686–3693.

[18] Sam Johnson and Mark Everingham, "Learning effective human pose estimation from inaccurate annotation," in *CVPR*, 2011, pp. 1465–1472.

[19] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *CVPR*, 2017, pp. 1077–1085.

[20] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian, "Pose-driven deep convolutional model for person re-identification," in *ICCV*, 2017, pp. 3980–3989.

[21] Ross Girshick, "Fast r-cnn," in *ICCV*, 2015, pp. 1440–1448.

[22] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al., "Spatial transformer networks," in *NIPS*, 2015, pp. 2017–2025.

[23] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM Multimedia*, 2014, pp. 675–678.

[24] Dangwei Li, Zhang Zhang, Xiaotang Chen, Haibin Ling, and Kaiqi Huang, "A richly annotated dataset for pedestrian attribute recognition," *arXiv preprint arXiv:1603.07054*, 2016.

[25] Yao Li, Guosheng Lin, Bohan Zhuang, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel, "Sequential person recognition in photo albums with a recurrent network," in *CVPR*, 2017, pp. 5660–5668.

[26] Feng Liu, Tao Xiang, Timothy M Hospedales, Wankou Yang, and Changyin Sun, "Semantic regularisation for recurrent image annotation," in *CVPR*, 2017, pp. 4160–4168.

[27] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016, pp. 2818–2826.