Grouping Attribute Recognition for Pedestrian with Joint Recurrent Learning *

Xin Zhao1, Liufang Sang1, Guiguang Ding1, Yuchen Guo1 Xiaoming Jin1
1Beijing National Research Center for Information Science and Technology(BNRist)
School of Software, Tsinghua University, Beijing 100084, China
{zhaoxin19,yuchen.w.guo}@gmail.com, slf12thuss@163.com, {dinggg,xmjin}@tsinghua.edu.cn

Abstract

Pedestrian attributes recognition is to predict attribute labels of pedestrian from surveillance images, which is a very challenging task for computer vision due to poor imaging quality and small training dataset. It is observed that semantic pedestrian attributes to be recognised tend to show semantic or visual spatial correlation. Attributes can be grouped by the correlation while previous works mostly ignore this phenomenon. Inspired by Recurrent Neural Network (RNN)’s super capability of learning context correlations, this paper proposes an end-to-end Grouping Recurrent Learning (GRL) model that takes advantage of the intra-group mutual exclusion and inter-group correlation to improve the performance of pedestrian attribute recognition. Our GRL method starts with the detection of precise body region via Body Region Proposal followed by feature extraction from detected regions. These features, along with the semantic groups, are fed into RNN for recurrent grouping attribute recognition, where intra group correlations can be learned. Extensive empirical evidence shows that our GRL model achieves state-of-the-art results, based on pedestrian attribute datasets, i.e. standard PETA and RAP datasets.

1 Introduction

Pedestrian attributes, e.g. age, gender, and clothing are humanly searchable semantic descriptions and can be used as soft-biometrics in visual surveillance applications such as person re-identification [Layne et al., 2012; Liu et al., 2012; Peng et al., 2016], face verification [Kumar et al., 2009], and human identification [Reid et al., 2014]. Attributes are robust against viewpoint changes and viewing condition diversity compared to low-level visual features. While pedestrian attribute recognition has been profitably attacked from a face recognition perspective, very few works focus on whole people body. There is inherently challenging to recognise pedestrian attributes from real-world surveillance images subject to the poor imaging quality and small training dataset. High imaging quality and large scale training data are not available for pedestrian attributes. For example, the two largest pedestrian attribute benchmark datasets PETA [Deng et al., 2014] and RAP [Li et al., 2016a] contain only 9500 and 33268 training images. Besides, recognising pedestrian attributes has to cope with images with poor quality, imbalance label and complex appearance variations in surveillance scenes.

Attribute recognition methods include hand-crafted feature methods, CNN methods and CNN-RNN methods. Early attribute recognition methods mainly rely on hand-crafted features like colour and texture [Layne et al., 2012; Liu et al., 2012; Jaha and Nixon, 2014]. Recently, deep learning based attribute models have been proposed due to the capacity to learn more expressive representations [Li et al., 2015; Fabbri et al., 2017; Liu et al., 2017b], which significantly improve the performance of pedestrian attribute recognition. For example, DeepMar method [Li et al., 2015] utilizes the prior knowledge in the object topology for attribute recognition and designs a weighted sigmoid cross-entropy loss to deal with the data imbalance problem whilst training attribute recognition model. Multi-directional attention modules are applied in an inception based deep model named HydraPlus Network [Liu et al., 2017b] to take the visual attention into consideration. CNN-RNN methods are proved to be a success in multi-label classification task to mine the dependency of labels [Li et al., 2017; Liu et al., 2017a]. A recurrent encoder-decoder framework is introduced into pedestrian attribute recognition task [Wang et al., 2017b], which aims to discover the interdependency and correlation among attributes with Long Short-Term Memory (LSTM) model.

Attributes of pedestrian always show semantic or visual spatial correlation by which they can be grouped. For example, BoldHair and BlackHair cannot occur on the same person while they are both related to the head-shoulders region of a person, so they can be in the same group to be recognised together. Existing methods try to mine the correlations of attributes separately but ignore both the intra-group semantic conflicts and the spatial neighborhood relationship of a group of attributes, which can actually improve the performance of pedestrian attribute recognition. There are two
types of semantic conflicts in the attribute prediction result. For example, a person cannot have the attribute *age 16-30* and *age 31-45* at the same time. If this occurs in the prediction result, it is termed as mutex co-occur. A person cannot be neither male or female. If this occurs, it is termed as none of any. Tab.1 shows the rate of semantic conflict of attributes *Age* and *Body Shape* with an existing pedestrian recognition method DeepMar [Li et al., 2015]. Moreover, the attributes are predicted separately with no attention to the spatial local attribute group, which makes it difficult to process spatial neighborhood relationship of attributes.

To address these problems, one idea is to take advantage of the interdependency and correlation among attributes [Chen et al., 2012; Li et al., 2015; Wang et al., 2016; 2017a; Zhu et al., 2017], while another idea focuses on particular spatial visual region for relevant attributes with intention to avoid the negative influence of the background [Li et al., 2016b; Liu et al., 2017b]. However, these two schemes are mostly studied independently in the existing methods.

In this work, we model both intra-group semantic mutual exclusion and inter-group correlations in an end-to-end recurrent architecture. A Grouping Recurrent Learning (GRL) framework is formulated to recognize pedestrian attributes by group step by step in order to pay attention to both the intra-group and inter-group relationship (including semantic and spatial). A novel grouping attribute recognition network is introduced which is specifically designed for sequential pedestrian attribute prediction by group. This RNN based model, which applies a sequential grouping attribute prediction, differs from the existing CNN based attribute prediction policy [Li et al., 2015; Fabbri et al., 2017; Liu et al., 2017b]. Moreover, it is an end-to-end single-model method with no need for preprocessing, compared to the multi-model Joint Recurrent Learning (JRL) method [Wang et al., 2017b]. More latent intra-group and inter-group dependency among grouped pedestrian attributes can be exploited, therefore the proposed method outperforms existing methods on the pedestrian attribute recognition task. In summary, we make the following contributions in this paper:

- We put forward a novel approach termed as GRL for pedestrian attribute recognition. To the best of our knowledge, it is the first work that predicts attributes group by group via mining both semantic and spatial correlations in attribute groups.
- A single-model end-to-end architecture, which is easier to train, is adopted without much more preprocessing prior to feature extraction and multi-model voting after attribute prediction.

- A recurrent learning method is proposed for mining the inter-group attribute correlations.

## 2 Related Work

### 2.1 Pedestrian Attribute Recognition

Semantic pedestrian attributes have been extensively exploited for person identification [Jaha and Nixon, 2014] and re-identification [Layne et al., 2012; Liu et al., 2012; Peng et al., 2016]. Attribute recognition methods include hand-crafted feature methods, CNN methods and CNN-RNN methods. Earlier methods typically model multiple attributes independently and train a separate classifier for each attribute based on hand-crafted features such as color and texture histograms [Layne et al., 2012; Liu et al., 2012; Jaha and Nixon, 2014]. Later on, inter-attribute correlation is considered as an extra information for improving prediction performance, e.g. graph model based methods to capture attribute co-occurrence likelihoods by using conditional random field or Markov random field [Chen et al., 2012; Deng et al., 2015; Shi et al., 2015]. But existing graph models are expensive to compute when dealing with a large set of attributes. Restricted to the poor discriminability of handcrafted features, these methods do not work well.

Recently, deep CNN based methods [Zhu et al., 2015; Li et al., 2015; Sudowe et al., 2015; Fabbri et al., 2017; Liu et al., 2017b] have been adopted in pedestrian attribute recognition task to learn more expressive representations which significantly improve the performance of pedestrian attribute recognition. DeepMar model [Li et al., 2015] utilizes the prior knowledge in the object topology for attribute recognition and designs a weighted sigmoid cross entropy loss to deal with the data imbalance problem while attribute recognition model training. Spatial attention methods [Liu et al., 2017b; Fabbri et al., 2017] are proposed to avoid the negative effect of irrelevant image region. Although the CNN based methods learn more more expressive pedestrian representations by using deep convolutional network, they are always insufficient in mining the correlations of attributes.

A CNN-RNN based encoder-decoder framework is proposed in [Wang et al., 2017b], which aims to discover the inter-dependency and correlation among attributes with LSTM model. However, semantic mutex constraint and the spatial neighborhood are not taken into account in this method. Additionally, predicting attributes one by one with multi-model voting afterwards is very expensive in computation.

### 2.2 Body Region Proposal

Body region proposal problem can be considered as an object detection problem. Region-based Convolutional Network (RCNN) methods are proposed in object detection task and achieve success [Girshick, 2015; Ren et al., 2015]. Fast R-CNN method is proposed for a fast object detection by sharing computation of convolutional feature map with ROI(Region of interest) pooling layer. And then a fully-convolutional object detection framework that simultaneously predicts object bounds and objectness scores at each position termed as Region Proposal Networks (RPN) [Ren et al., 2015].


<table>
<thead>
<tr>
<th>Attribute</th>
<th>Age</th>
<th>Body Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mutex Co-occur</td>
<td>158</td>
<td>562</td>
</tr>
<tr>
<td>None of Any</td>
<td>134</td>
<td>0</td>
</tr>
<tr>
<td>Total Error Number</td>
<td>292</td>
<td>562</td>
</tr>
<tr>
<td>Total Image Number</td>
<td>8317</td>
<td>8188</td>
</tr>
<tr>
<td>Semantic Error Rate</td>
<td>3.51%</td>
<td>6.86%</td>
</tr>
</tbody>
</table>

Table 1: The rate of semantic conflict of *Age* and *Body Shape* in RAP test set in the prediction result of DeepMar.
...are connected together with a batch normalization layer (BN) afterwards. The batch normalization layer first

Long range dependency of input sequence can be captured by

3.2 Long Short-Term Memory(LSTM) Model

x step t, the LSTM updates using the input

θ ϕ where

3.1 Recurrent Neural Network

3 Background

Body region feature extraction plays an important role in
distinguishing individuals. Local details can be better de-
dscribed by the region features. A complete body region pro-
posal consists of two steps, which are body joint localization
and body region proposal. RPN is introduced for body region
proposal [Zhao et al., 2017] in a person re-identification task for
more accuracy local features, where a fully convolutional
network (FCN) is adopted to predict the localization of body
joints, and the position of joints is used for body region gener-
ation. In this work, we use this body region proposal method
to detect body parts in pedestrian images and use the relevant
spatial region for grouping attribute recognition. Fig.1 shows
the pipeline of body region proposal.

3 Background

3.1 Recurrent Neural Network

RNN is a neural network consisting of an internal hidden state
h ∈ Rd and operating on a variable-length input sequence
X = (x1, x2, . . . , xt, . . . ). At each time step t, the RNN
takes sequentially an element xt of X and then updates its
hidden state ht as:

where φθ denotes the non-linear activation function parameter-
sed by θ.

3.2 Long Short-Term Memory(LSTM) Model

Long range dependency of input sequence can be captured by
LSTM [Hochreiter and Schmidhuber, 1997] as recurrent neu-
ron for sequential grouping attribute prediction. LSTM is also
effective to handle the common gradient vanishing and explod-
ing problems in training RNN. Particularly, at each time
step t, the LSTM updates using the input xt and the LSTM
previous status h_{t-1} ∈ Rd, and c_{t-1} ∈ Rd as:

where sigmoid(·) refers to the logistic sigmoid function,
tanh(·) refers to the hyperbolic tangent function and the
operator ⊙ refers to the element-wise vector product. The
LSTM contains four multiplicative gating units: forget gate
f ∈ Rd, input gate i ∈ Rd, output gate o ∈ Rd, input mod-
ulation gate g ∈ Rd, with corresponding matrix and bias pa-
rameters to be learned. The memory cell c_t depends on the
previous memory cell modulated by the forget gate and the
the current input. Therefore, LSTM learns to forget its pre-
vious memory and exploit its current input selectively. And
the output gate o learns how to transfer the memory cell c_t to
the hidden state h_t. These gates learn to effectively modulate
the behaviour of input signal propagation through the recur-
rent hidden states in order to capture long-term dependency
in sequence data.

4 Grouping Joint Recurrent Learning for Pedestrian Attribute Recognition

4.1 Problem Definition

Grouping pedestrian attribute recognition can be defined as
follows. We are given n training images {I1, . . . , In} and
each image Im has km visual attribute tags. Each visual at-
tribute tag belongs to set T = {T1, . . . , Tk_T}, where K_T is
the size of T. And G = {G1, . . . , GK_C} is a set partitioning
of T, where Gi ∩ Gj = ø(i ≠ j) and all combinations of Gi
is the entire set T. Tags in the same group are with semantic
or spatial constraint with each other. For each image there is
a label vector y_m ∈ {0, 1}^k_T where y_{m,j} = 1 if Im has tag
Tj and y_{m,j} = 0 otherwise. We aim to learn attribute recog-
nition models R^T: I → {0, 1}^k_T to recognise the attributes
of image Im.

4.2 Network Architecture

The network architecture is shown in Fig.2. For each pedes-
trian image Im, we use a fully convolutional network to de-
tect the joints of the body. And then we use a body region
propositional network to generate the head, upper body and lower
body region of this person. Im is fed into an inception based
CNN model with the results of body region proposal for ROI
average pooling. The ROI average pooling operation can ex-
tract the features in the particular region from the feature map
extracted from inception module, so that we can take advan-
tage of the spatial neighborhood of relevant group more eas-
ily. For example, the hair style, glasses and hat are all in the
head region, they are in the same group and these attributes
are predicted at the same time. There are attributes related to
the whole body region with semantic relationships, by which
they can be grouped into several groups in order to take ad-
vantage of the semantic correlation with each other.

As is shown in Fig.2, all attributes in the same group share
the same fully connected feature, and the features of all the
groups are fed into a LSTM unit for recurrent grouping at-
prediction. Each output of LSTM is fully connected into a predic-
tion vector which has the same dimension as the number of attributes in the relevant group. The prediction
vectors are connected together with a batch normalization
layer (BN) afterwards. The batch normalization layer first
normalizes the prediction vector into a vector with zero mean and unit variance, and then scales it and adds a bias in.

The batch normalize layer is used to balance the positive and negative outputs of this network. The output of batch norm layer is used to compute the weighted sigmoid cross-entropy loss, which will be stated in Section 4.3.

4.3 Loss Function and Optimization

The sigmoid cross entropy loss, which is defined in Eq.3, is introduced in multi-class classification problem.

\[
Loss = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K_F} y_{ij} \log(p_{ij})
+ (1 - y_{ij}) \log(1 - \hat{p}_{ij})
\]

\[
\hat{p}_{ij} = \frac{1}{1 + \exp(-x_{ij})}
\]

where \(p_{ij}\) is the output probability for the jth attribute of example \(I_i\), \(y_{ij}\) is the ground truth label which represents whether \(I_j\) has the jth attribute or not, \(x_i\) is the output of network fed with \(I_i\).

As is stated in [Li et al., 2015], attributes do not always have uniform distribution, and sometimes it is more like an unbalanced distribution, especially in surveillance scenarios. So we use the weighted sigmoid cross-entropy loss proposed in [Li et al., 2015] as follow to solve this problem.

\[
Loss = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K_F} w_j y_{ij} \log(p_{ij})
+ (1 - y_{ij}) \log(1 - \hat{p}_{ij})
\]

\[
w_j = \exp(p_j)
\]

where \(p_j\) is the positive ratio of jth attribute in the training set. \(w_j\) donates the learning weights of positive samples to deal with the imbalance label. We train the attribute recognition model with Stochastic Gradient Descent (SGD) algorithm.

<table>
<thead>
<tr>
<th>Group</th>
<th>Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>male or female</td>
</tr>
<tr>
<td>Age</td>
<td>age 16-30, age 31-45, age 46-60, age 60+</td>
</tr>
<tr>
<td>Head</td>
<td>hair length, muller, hat, glasses</td>
</tr>
<tr>
<td>Upper Body</td>
<td>clothes style, logo, casual or formal</td>
</tr>
<tr>
<td>Lower Body</td>
<td>clothes style, casual or formal</td>
</tr>
<tr>
<td>Footware</td>
<td>footware style</td>
</tr>
<tr>
<td>Accessories</td>
<td>backpack, messenger bag, plastic bag etc</td>
</tr>
</tbody>
</table>

Table 2: The groups of 35 binary attributes in PETA dataset

<table>
<thead>
<tr>
<th>Group</th>
<th>Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>male or female</td>
</tr>
<tr>
<td>Age</td>
<td>age 16-30, age 31-45, age 45+, gender</td>
</tr>
<tr>
<td>Body Shape</td>
<td>slightly fat, standard, slightly thin</td>
</tr>
<tr>
<td>Role</td>
<td>customer, uniform</td>
</tr>
<tr>
<td>Head</td>
<td>hair style, hair color, hat, glasses</td>
</tr>
<tr>
<td>Upper Body</td>
<td>clothes style, clothes color</td>
</tr>
<tr>
<td>Lower Body</td>
<td>clothes style, clothes color</td>
</tr>
<tr>
<td>Footware</td>
<td>footware style, footware color</td>
</tr>
<tr>
<td>Accessories</td>
<td>backpack, single shoulder bag, handbag etc</td>
</tr>
<tr>
<td>Action</td>
<td>telephoning, gathering, talking, pushing etc</td>
</tr>
</tbody>
</table>

Table 3: The groups of 51 binary attributes in RAP dataset

5 Experiment

5.1 Datasets

For evaluations, we used the two largest publicly available pedestrian attribute datasets: (1) The PE redesTrain Attribute (PETA) [Deng et al., 2014] dataset consists of 19000 person images collected from 10 small-scale person datasets. Each image is labelled with 65 attributes (61 binary + 4 multi-valued). Following the same protocol as [Deng et al., 2015; Li et al., 2015], we divide the whole dataset into three non-overlapping partitions: 9500 for model training, 1900 for verification, and 7600 for model evaluation. And we select 35 attributes from PETA dataset in our experiments. (2) The Richly Annotated Pedestrian (RAP) attribute dataset [Li et al., 2016a] has 41585 images drawn from 26 indoor surveillance cameras. Each image is labelled with 72 attributes (69
Table 4: Evaluation on PETA and RAP with bold best result and underline second best result. The first group is CNN method with small model such as Alexnet, while the second group is based on larger CNN model (Inception-v3 or Resnet50). The third group is CNN-RNN joint learning method. All above are single model methods, while JRL* uses multi-model ensemble.

5.2 Evaluation

Metrics. We use four metrics to evaluate attribute recognition performance. (1) Class-centric: For each attribute tag, we compute the classification accuracy of positive and negative samples respectively, average them to obtain a mean of accuracy termed as mA. (2) Instance-centric: For each instance, we measure the attribute prediction precision and recall as well as the F1 score based on precision and recall.

Competitors. Our method is compared against 8 state-of-the-art methods including 5 CNN based deep learning attribute recognition methods and 3 CNN-RNN based joint learning models. Attributes Convolutional Network (ACN) [Sudowe et al., 2015] trains jointly a CNN model for all attributes, and sharing weights and transfer knowledge among different attributes. DeepSAR [Li et al., 2015] is a deep model that processes attribute classes individually by training multiple attribute-specific models based on Alexnet. Different with DeepSAR [Li et al., 2015], DeepMar [Li et al., 2015] considers additionally inter-attribute correlation by learning all attributes in a single model. We train an inception based DeepMar model for fair comparison. HydraPlus-Net [Liu et al., 2017b] is an inception based multi-directional attention network to capture the spatial information of local attribute for better recognition performance. Resnet based Generative Adversarial Models are adopted in pedestrian attribute recognition to improve the accuracy of recognition termed as Generative Adversarial Pedestrian Attribute Recognition (GAPAR) [Fabbri et al., 2017] in this work. Contextual CNN-RNN (CTX CNN-RNN) [Li et al., 2017] is a CNN-RNN based sequential prediction model designed to encode the scene context and inter-person social relations for modeling multiple people in an image. Semantically Regularised CNN-RNN (SR CNN-RNN) [Liu et al., 2017a] is a state-of-the-art multi-label image classification model that exploits the ground truth attribute labels for strongly supervised deep learning and richer image embedding. Multi-model Joint Recurrent Learning (JRL) [Wang et al., 2017b] method is proposed for pedestrian attribute recognition which introduce an encoder-decoder architecture to process image context and attribute correlation.

Implementation Details. Our model is trained with tensorflow. And it is finetuned from the Inception-v3 model pre-trained from ImageNet image classification task. The body region proposal network is trained with MPII human pose dataset [Andriluka et al., 2014] as the model stated in [Zhao et al., 2017]. The optimization algorithm used in training the proposed model is SGD. The initial learning rate of training is 0.1 and reduced to 0.001 by a factor of 0.1 at last.

Results. The experiment results of our method and competitors are in Tab.4. The methods in Tab.4 are divided into 4 groups, which are CNN small model, CNN large model, CNN-RNN model and multi-model method. Our GRL method outperforms the SR CNN-RNN [Liu et al., 2017a] which is the state-of-the-art single model CNN-RNN method in all the four metrics improving 3.87% and 3.86% in mA and F1 of PETA dataset, while the numbers in RAP dataset are 6.99% and 3.46%. And GRL outperforms JRL* [Wang et al., 2017b] in mA and F1 score although JRL* is a multi-model ensemble method, improving 1.03% and 1.09% in PETA as well as 3.39% and 0.71% in RAP. Compared to large model CNN method, GRL is better than GAPAR [Fabbri et al., 2017] method in all the metrics in RAP improving 1.47% in mA where GAPAR is better than other methods. The DeepMar [Li et al., 2015] based on Inception-v3 is better in the instance-centric metric than class-centric metric. GRL also achieves little advantage in instance-centric F1 score (0.83% in PETA and 0.99% in RAP). The experiment result shows clearly the benefit of the proposed GRL approach in pedestrian attribute recognition. This is mainly due to the capacity of GRL in mining both the intra-group and inter-group correlations.
Table 5: The experiment result of Full GRL system and GRL system without ROI pooling from Body Region Proposal compared to the DeepMar (Inception-v3) method as baseline. The best result is in bold.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>mA</th>
<th>precision</th>
<th>recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>PETA</td>
<td>Baseline</td>
<td>81.50</td>
<td>89.70</td>
<td>81.90</td>
<td>85.68</td>
</tr>
<tr>
<td></td>
<td>GRL(no ROI)</td>
<td>85.66</td>
<td>84.98</td>
<td>87.20</td>
<td>86.01</td>
</tr>
<tr>
<td></td>
<td>GRL</td>
<td>86.70</td>
<td>84.34</td>
<td>88.82</td>
<td>86.51</td>
</tr>
<tr>
<td>RAP</td>
<td>Baseline</td>
<td>76.10</td>
<td>82.20</td>
<td>74.80</td>
<td>78.30</td>
</tr>
<tr>
<td></td>
<td>GRL(no ROI)</td>
<td>79.68</td>
<td>78.60</td>
<td>78.90</td>
<td>78.70</td>
</tr>
<tr>
<td></td>
<td>GRL</td>
<td>81.20</td>
<td>77.70</td>
<td>80.90</td>
<td>79.29</td>
</tr>
</tbody>
</table>

Table 6: The experiment result of logical optimized prediction order (global to local) compared to random prediction order of GRL. The best result is in bold.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Metric</th>
<th>Method</th>
<th>mA</th>
<th>precision</th>
<th>recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>PETA</td>
<td>random order</td>
<td>global to local</td>
<td>85.81</td>
<td>84.32</td>
<td>87.49</td>
<td>85.89</td>
</tr>
<tr>
<td></td>
<td>random order</td>
<td>global to local</td>
<td>86.70</td>
<td>84.34</td>
<td>88.82</td>
<td>86.51</td>
</tr>
<tr>
<td>RAP</td>
<td>random order</td>
<td>global to local</td>
<td>80.02</td>
<td>77.99</td>
<td>79.21</td>
<td>78.70</td>
</tr>
<tr>
<td></td>
<td>random order</td>
<td>global to local</td>
<td>81.20</td>
<td>77.70</td>
<td>80.90</td>
<td>79.29</td>
</tr>
</tbody>
</table>

5.3 Further Analysis and Discussions

Effect of Body Region Proposal and Grouping Recurrent Recognition. The improvement of GRL method comes from two aspects which are spatial attention and semantic correlation mining. In this section, we will discuss how much improvement this two aspects bring. First of all, we do not make use of the information of Body Region Proposal. The full body features from the CNN are fed into the LSTM unit directly without ROI average pooling. Compared to the baseline method, which is an inception based DeepMar model, this method only uses Grouping Recurrent Recognition. So the improvement from the baseline donates the effect of Grouping Recurrent Recognition. Compared to the full GRL system, this method is lack of the spatial attention from Body Region Proposal. So the difference from the full GRL system donates the effect of Body Region Proposal. The experiment results are listed in Tab.5.

From Tab.5 we can see that GRL without ROI achieves a gain of 4.16% in mA and 0.33% in F1 score on PETA from baseline, as well as 3.58% and 0.40% on RAP. The full system improves 1.04% in mA and 0.50% in F1 score, while the numbers are 1.52% and 0.59% on RAP. We can see that both of the two components of GRL make sense and GRL achieves more improvement in mA from Grouping Recurrent Recognition than Body Region Proposal.

Effect of Prediction Order of LSTM. The prediction order is an important influence factor for the recognition accuracy because the attributes to be recognised at the very beginning cannot observe much more relevant recognition result. So we should put the global attributes which can be easily recognised without relying badly on others first. For example, gender and age can easily be recognised even though many other attributes are not clear. The recognition of gender is very helpful in predicting some other related attributes (e.g. clothing and footware). In this section, we show the experiment result of logical optimized prediction order from group to local group and that of a random order which are listed in Tab.6. The logical optimized prediction order in PETA and RAP is as the order in Tab.2 and Tab.3.

The experiment result listed in Tab.6 confirms our inference that the logical optimized order is better than a random one, which brings an improvement of 0.89% and 0.62% as well as 1.08% and 0.59% in mA and F1 of both datasets.

The effect of attribute correlations are examined more carefully on the GRL model performance. Fig.3 shows two examples from RAP dataset for qualitative analysis which indicates that a proper prediction order is necessary for grouping pedestrian attribute recognition. Non-sequence prediction model DeepMar misses hair long and gets wrong prediction of the age for the left person. And it also misses hair long and cotton clothes for the right one. In contrast, our GRL method gets the right prediction perfectly when the prediction is in a right order. For example, if the model has come to a conclusion that a person is female, the chance she is with long hair is higher. So when predicting the attribute group related to hair after getting gender information, the attribute hair long will be rightly recognised. The prediction for some global attributes (e.g. age and body shape) is not so much related to the prediction result of other local attributes, so that they should be determined by the overall vision features as much as possible to avoid the misleading of wrong local attribute tags. Attributes like that should be predicted at the beginning of the sequence. GRL gets wrong body shape and age results when predicting in a wrong order as is shown in Fig.3.

6 Conclusion

In this work, we present a novel end-to-end deep Grouping Recurrent Learning (GRL) model for exploring the intra-group relationship including semantic dependency and spatial neighborhood as well as inter-group pedestrian attribute correlations. Spatial attention from Body Region Proposal and Grouping Recurrent Recognition are adopted in the GRL architecture. Our GRL model outperforms a wide range of existing pedestrian attribute recognition methods. Extensive experiments demonstrate the advantages of spatial attention from Body Region Proposal and Grouping Recurrent Recognition on two pedestrian benchmarks. Moreover, a logical optimized prediction order is proved to lead to a better result.
References