DEEP PEDESTRIAN ATTRIBUTE RECOGNITION BASED ON LSTM

Zhong Ji, Weixiong Zheng, Yanwei Pang

School of Electrical and Information Engineering
Tianjin University, Tianjin 300072, China
Email: jizhong@tju.edu.cn, zhengweixiong1203@gmail.com, pyw@tju.edu.cn

ABSTRACT

Automatically recognizing attributes such as gender, age, footwear and clothing style from pedestrian images at far distance is an important task in surveillance scenarios. However, the appearance diversity and ambiguity in these images make it a challenging task. This paper presents an end-to-end Neural Pedestrian Attribute Recognition (Neural PAR) model to address these challenges. Rather than taking it as a recognition problem like previous methods, Neural PAR formulates it as an end-to-end image to attribute description problem. To this end, the training images and their corresponding attributes are used as inputs. Specifically, the attributes are concatenated into different attribute descriptions to well contextualize the potential relationships among them. Then, a neural network model is trained based on CNN and LSTM to learn the complex relations between visual features and their corresponding attributes. Extensive experiments show that the proposed Neural PAR significantly outperforms the state-of-the-art methods on the benchmark PETA dataset.

Index Terms—Pedestrian attribute recognition, LSTM, Visual surveillance, Attribute prediction.

1. INTRODUCTION

Pedestrian attributes recognition is an emerging research topic in intelligent surveillance domain. It attracts increasing attention recently because of its practical applications in person retrieval [1, 2], identification [3, 4] and re-identification [5, 6]. In many real-world surveillance scenarios, close-shots of body and face are hard to obtain. As a result, the full body appearance in far-view is the main visual information for predicting pedestrian attributes.

However, there are two fundamental challenges in this task [7, 8]. The first one is appearance diversity. There exist large intra-class variations for the same attribute due to the diverse pedestrian appearances, such as different hats, clothes and backpacks. Moreover, different illuminations and camera angles make it more diverse. The other one is appearance ambiguity. Since images are captured at far distance, the image resolution is quite low. As illustrated in Fig. 1, some attributes are even hard to recognize by humans.

Fig. 1. Sample images in PETA dataset.

To address these challenges, some studies have been proposed in recent years. One direction focused on using independent attribute classifiers to recognize attributes. For example, to train classifiers for every binary attribute. Zhu *et al.* [7] and Deng *et al*. [8] used algorithms of AdaBoost and SVM to train independent attribute classifiers, respectively. Further, [9] explored the hidden neighborhood information of multiple pedestrian images by using Markov Random Field (MRF) and combined foreground segments with background context for improving pedestrian representation. However, these methods have some limitations because independent classifiers cannot well take advantage of the relationship among attributes. To overcome this drawback, recent direction starts to exploit the relationship among attributes. For example, Chen *et al*. [10] applied conditional

This work was supported by the National Natural Science Foundation of China under Grants 61472273 and 61632018.

random field (CRF) to find the mutual dependencies by using the SVM margins from the independently trained attribute classifiers. And Li *et al.* [11] proposed a DeepMAR framework based on Convolutional Neural Networks (CNN) to recognize multiple attribute simultaneously with a loss function jointly considering all attributes. These methods have more robust performance and build a balance between the independent decision score and attribute interactions.

Meanwhile, image caption technique has achieved great success in recent years mainly due to the great contextual modeling power of Recurrent Neural Network (RNN) and Long Short Term Memory (LSTM) [12, 13, 14]. The goal of image caption is to generate a sentence or sentences to accurately describe the content of an arbitrary image. Its general pipeline at training stage is to apply the image and its corresponding descriptions as the inputs, and then use the CNN-RNN architecture to learn the visual-semantic mapping and their contextual relations. Inspired by the idea of image caption, we treat the attribute recognition problem as an image caption task by randomly connecting the labeled attribute words as different sentences. In this way, the attribute sentences can be viewed as special image descriptions, and the predicted image caption is a combination of the image attributes. We name the proposed method as Neural PAR (Neural Pedestrian Attribute Recognition), and its flowchart is illustrated in Fig. 2.

It is worthwhile to highlight several aspects of the proposed Neural PAR approach here: 1) To the best of our knowledge, it is the first end-to-end approach using CNN-RNN architecture for pedestrian attribute recognition. 2） Different from existing methods that taking pedestrian attribute recognition as a recognition task, Neural PAR formulates it as an image to attribute description problem. 3) Extensive experiments on the benchmark dataset PETA demonstrate the superiority of Neural PAR against state-of-the-art methods.

2. DEEP PEDESTRIAN ATTRIBUTE RECOGNITION BASED ON LSTM

The flowchart of Neural PAR is shown in Fig. 2, which has mainly three parts: inputs, neural network model, and outputs, which will be described in detail in this section.

2.1. Inputs

The part enclosed by the green dotted line in Fig. 2 is the input at the training stage. There are two inputs, one is the pedestrian images, and the other is their corresponding labeled attribute sentences.

As we know, many pedestrian attributes are interrelated. For example, the feature of wearing skirt has a higher probability for women than men, men are more likely than women to have short hair. Thus the dressing and the hair length could be utilized to recognize the gender. To

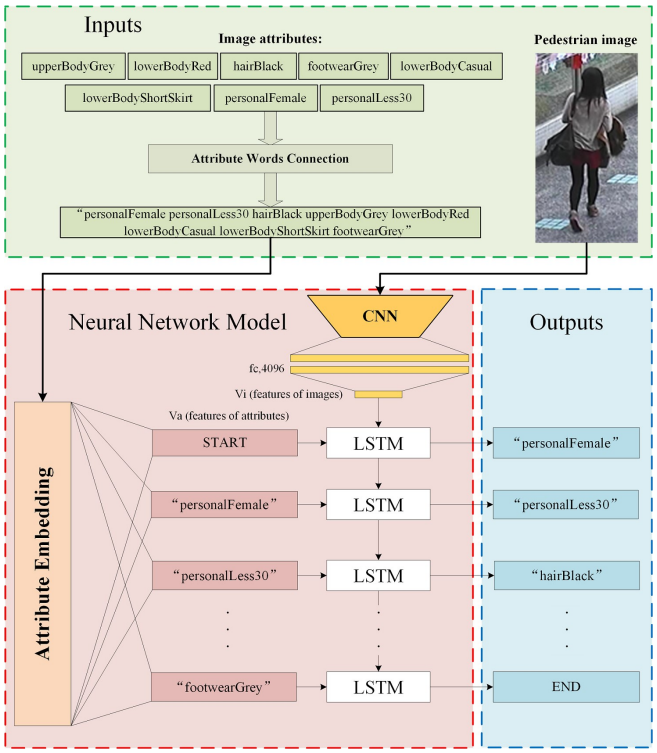

Fig. 2. Flowchart of the proposed Neural PAR.

well capture these relations, we consider attributes as words with semantic information so that semantic relevance of attributes can be utilized. To explore the semantic information of these attribute words, we spliced the original attribute labels into a serialized attribute label like a sentence describing the image (see Fig. 2). For example, "accessoryHat", "hairShort" and "personalMale" are original labels for a pedestrian image. After the attribute words connection, we get a new sentence label "accessoryHat hairShort personalMale" to describe the pedestrian image. Attribute words are assembled into an attribute description so that the contextual information of the sentence can be used to predict the attributes. We call this attribute description "attribute sentence". Therefore, the issue of pedestrian attribute classification is converted into a pedestrian attribute generation task. By randomly connecting the attribute words, we obtain several attribute sentences for a given pedestrian image.

2.2. Neural Network Model

The part enclosed by the red dotted line in Fig. 2 is the Neural Network Model. It includes three parts: 1) Attribute embedding layer, 2) CNN part, and 3) Multimodal LSTM part.

Attribute embedding: This layer aims at extracting the attribute vectors and then transforming them into a common space. First, to represent the attribute sentences with feature

vectors, we build an attribute dictionary of size $N+1$, where N represents the number of the attribute category that appears in the training set. One is added on N because there is a "START" token to control the generation of attribute description. We represent each attribute and special token as a one-hot vector V_d with dimensionality equal to the size of the attribute dictionary. Thus the final text input to the neural network model are $M+1$ vectors, where M is the number of attributes in one attribute sentence.

Then, to transform these attribute vectors into a common space, we build a lookup-table W_a of size $L\times(N+1)$ with parameters to be learned. The output of the lookup-table is the representation of attribute vector V_a:

$$V_a = W_a V_d \tag{1}$$

Representing images with CNN: We use a CNN to extract the images' visual information. Particularly, we use the model of VGG-16 net [15] pre-trained on ImageNet. Additionally, we add a fully connected layer to transform the feature vector dimensionality from 4096 to L:

$$V_i = CNN(I_i) \tag{2}$$

where I_i is the i-th input image, and V_i is its visual feature in the common space.

Multimodal LSTM: After extracting the visual feature V_i and the attribute vector V_a, we require a model to learn the mapping between the visual features and the attribute vectors. Moreover, a challenge is that the model should generate variable-sized attributes description since the number of pedestrian attributes is often uncertain. To this end, we borrow the RNN framework in image caption to realize it. Specifically, we define $h_{(t)}$ as a memory or hidden state of LSTM at time step t. When given a new input $x_{(t)}$, a non-linear function f will act on $x_{(t)}$ and $h_{(t)}$ to update $h_{(t)}$ into $h_{(t+1)}$:

$$h_{(t+1)} = f(x_{(t)}, h_{(t)}) \tag{3}$$

We build one LSTM layer as the function f in Neural PAR. As a special kind of RNN, LSTM is explicitly designed to avoid the long-term dependency problem of RNN, which means it can remember information for long periods of time. It has been widely used in sequence tasks [16, 17] and achieved great success.

Modal training: Let's define $V_{l(t)}$ the target label at time step t, which is represented by one-hot vector of size $N+1$ (There is an additional "END" token added on N). Therefore, $V_{l(t)}$ is got by a conditional probability $p(V_{l(t)} \mid V_i, V_{a(0)}, ..., V_{a(t-1)})$. Particularly, the last target attribute is the "END" token when given the last attribute of attribute description. In this way, the Neural PAR is trained to predict attributes one by one when given an image. As shown in Fig. 2, we unroll the LSTM by time steps to clearly illustrate the process of Neural PAR:

$$h_{(t+1)} = LSTM(V_{a(t)}, V_i, h_{(t)}), t = 0 \tag{4}$$

$$h_{(t+1)} = LSTM(V_{a(t)}, h_{(t)}), t \in \{1...M\} \tag{5}$$

$$p_{(t+1)} = softmax(h_{(t)}), t \in \{0...M\} \tag{6}$$

where M is the attribute number of an image, V_i is the visual vector of i-th image extracted by CNN and $V_{a(t)}$ is the t-th attribute word vector of attribute description at time step t, p is the output of a Softmax function which represents the probability of belonging to each attribute. The cost function is the sum of the negative log likelihood of the correct attributes assigned to the target labels at each time step:

$$C(V_i, V_a) = -\sum_{t=0}^{M} \log p_{(t)}(V_{l(t)}) \tag{7}$$

We minimize this cost function to learn the parameters in LSTM, CNN, and the embedding layer.

2.3. Outputs

The part enclosed by the blue dotted line in Fig. 2 is the outputs of Neural PAR. Similar to image caption in which the generated sentences conform to grammatical rules, our proposed Neural PAR can also generate attributes in a specified order we define. This is due to the powerful learning ability of LSTM. As we can see from Fig. 2, attribute words are generated one after another in the same order as they input (for example, from whole appearance of the pedestrian to attributes of head, upper body, lower body and feet).

2.4. Attributes Generation at testing time

At the testing time, our Neural PAR generates an attribute sentence for a given pedestrian image. After the representation V_i of image I is computed, we send the "START" vector V_s and V_i to LSTM to generate the first attribute A_1 by selecting the one with highest probability. At next step, the attribute vector V_{a1} of A_1 is used as input of LSTM to generate the second attribute A_2. The rest can be done in the same manner until the "END" token is generated.

3. EXPERIMENTS

3.1. Dataset and experimental settings

Comprehensive experiments on PEdesTrian Attribute (PETA) [8] are conducted to demonstrate the superiority of Neural PAR, since PETA is one of the biggest challenging pedestrian attribute datasets used for benchmark evaluation

to date. It consists of 19,000 pedestrian images which includes 10 subsets such as PRID [18], CUHK [19] and VIPeR [20]. Because of the differences among subsets, pedestrian images in PETA have different resolutions, viewpoints, scenes and illuminations. They are labeled with 61 binary attributes and 4 multi-class attributes. Some images in PETA has been shown in Fig. 1.

We follow the same setting of many other methods [8, 9, 11], where the images in PETA are randomly divided into 9,500 for training, 1,900 for verification and 7,600 for testing. Each image is resized into 224×224 pixels for adapting to the VGG-16 net. And the VGG-16 net is fine-tuned with the training images. Five attribute sentences for a given training image are applied for inputs. As for LSTM, the input vector dimensionality of L is set to 512, and the number of hidden nodes in the layer is also set to 512. The attribute dictionary size N is 105 because we use all the attributes in the training set to construct the dictionary. The model is optimized by stochastic gradient descent.

3.2. Experimental results

Three state-of-the-art methods are chosen for comparison. 1) ikSVM [8], using features of 8 color channels to train an SVM classifier. 2) MRFr2 [9], forming multiple pedestrian images as a Markov Random Field (MRF) graph to exploit the hidden neighborhood information and using extracted foreground segments of pedestrian and the whole image to improve pedestrian representation. 3) DeepMAR [11], a CNN-based model that constructs a loss function considering all attributes together.

For fair comparison, we select the same 35 attributes in [8, 9, 11] for evaluation. The comparative results are illustrated in Table 1. From the results we can find that Neural PAR achieves a quite good average accuracy of 90.8%, which is clearly outperforms all the comparative approaches. The absolute performance gains for average accuracy of Neural PAR against DeepMAR, MRFr2 and ikSVM are 8.2%, 15.2%, and 21.3%, respectively. Specifically, Neural PAR achieves the best performance on 28 attributes among all the 35 attributes, such as "Age46-60", "AgeAbove60", "Backpack", and "CarryingOther".

Additionally, we also explore the influence of attribute sentence number. Multiple attributes with different orders offer different contextual information to LSTM layer, which may lead to a better performance. For example, when only one attribute sentence for a given training image is applied for inputs, the average accuracy is 88.1%, which is inferior to that with five attribute sentences in 2.7% absolute gains. However, the performance is steady when the number is larger than 5. The reason behind it may be that five sentences are enough to offer comprehensive contextual information. More attribute sentences do not help improving the performance.

Table 1: Attributes recognition accuracy (%) comparison on PETA dataset. Bold font denotes the better case. Ratio is the proportion of positive sample attributes in all samples.

Attribute	ikSVM [8]	MRFr2 [9]	DeepMAR [11]	Neural PAR (Proposed)	Ratio
Age16-30	80.4	**86.8**	85.8	83.8	0.497
Age31-45	73.6	**83.1**	81.8	81.2	0.329
Age46-60	73.1	80.1	86.3	**94.6**	0.102
AgeAbove60	87.2	93.8	94.8	**98.7**	0.062
Backpack	66.7	70.5	82.6	**86.4**	0.197
CarryingOther	64.6	73.0	77.3	**84.9**	0.199
Casual lower	70.7	78.2	84.9	**92.7**	0.861
Casual upper	70.3	78.1	84.4	**92.1**	0.853
Formal lower	71.0	79.0	85.2	**92.9**	0.138
Formal upper	70.0	78.7	85.1	**93.3**	0.134
Hat	82.3	90.4	91.8	**96.8**	0.102
Jacket	67.7	72.2	79.2	**92.9**	0.069
Jeans	74.9	81.0	**85.7**	84.6	0.306
Leather Shoes	78.9	87.2	**87.3**	87.2	0.296
Logo	51.1	52.7	68.4	**94.6**	0.04
Long hair	71.5	80.1	88.9	**91.6**	0.238
Male	79.7	86.5	89.9	**90.6**	0.549
MessengerBag	71.8	78.3	82.0	**83.1**	0.296
Muffler	88.0	93.7	96.1	**98.4**	0.084
No accessory	76.8	82.7	85.8	**88.4**	0.749
No carrying	70.4	76.5	83.1	**84.0**	0.276
Plaid	64.0	65.2	81.1	**98.1**	0.027
Plastic bag	74.9	81.3	87.0	**94.9**	0.077
Sandals	50.3	52.2	67.3	**97.2**	0.02
Shoes	70.6	78.4	**80.0**	76.7	0.363
Shorts	56.0	65.2	80.4	**97.3**	0.035
ShortSleeve	71.3	75.8	87.5	**91.0**	0.142
Skirt	64.0	69.6	82.2	**96.5**	0.046
Sneaker	67.5	75.0	78.7	**83.9**	0.216
Stripes	51.5	51.9	66.5	**98.1**	0.017
Sunglasses	52.4	53.5	69.9	**96.7**	0.029
Trousers	74.0	82.2	**84.3**	79.8	0.515
Tshirt	64.3	71.4	83.0	**93.0**	0.084
UpperOther	80.7	**87.3**	86.1	83.9	0.456
V-Neck	51.1	53.3	69.8	**98.6**	0.012
AVERAGE	69.5	75.6	82.6	**90.8**	*

4. CONCLUSION

A novel end-to-end approach with CNN-RNN architecture for pedestrian attribute recognition has been presented in this paper. The inputs are the pedestrian images and their corresponding attribute sentences, which are the connections of each attribute. Experimental results on PETA have well demonstrated the superiority of the proposed Neural PAR, and also proved the effectiveness of treating attribute recognition as an attribute generation or description problem.

5. REFERENCES

[1] A. Dantcheva, A. Singh and P. Elia, "Search pruning in video surveillance systems: Efficiency-reliability tradeoff," *2011 IEEE International Conference on Computer Vision Workshop*, pp. 1356-1363, 2011.

[2] J. Emad Sami, and M. Nixon, "Analysing Soft clothing biometrics for retrieval," *International Workshop on Biometric Authentication*, pp.234-245, 2014.

[3] E. Martinson, W. Lawson and J. G. Trafton, "Identifying people with soft-biometrics at fleet week," *IEEE International Conference on Human-Robot Interaction*, pp. 49-56, 2013

[4] D. Reid, M. Nixon and S. Stevenage, "Soft biometrics; human identification using comparative descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6), pp.1216-1228, 2014.

[5] X. Liu, M. Song, Q. Zhao, D. Tao, C. Chen, and J. Bu, "Attribute restricted latent topic model for person re-identification," *Pattern recognition*, 45(12), pp.4204-4213, 2012.

[6] A. Li, L. Liu, K. Wang, S. Liu, and S. Yan, "Clothing attributes assisted person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, 25(5), pp. 869-878, 2015.

[7] J. Zhu, S. Liao, Z. Lei, D. Yi, and S. Z. Li, "Pedestrian attribute classification in surveillance: database and evaluation," *IEEE International Conference on Computer Vision Workshops*, pp. 331-338, 2013.

[8] Y. Deng, P. Luo and C. Loy, "Pedestrian attribute recognition at far distance," *ACM International Conference on Multimedia*, pp. 789-792, 2014.

[9] Y. Deng, P. Luo and C. Loy, "Learning to recognize pedestrian attribute," *arXiv preprint arXiv:1501.00901*, 2015.

[10] H. Chen, A. Gallagher and B. Girod, "Describing clothing by semantic attributes," *European Conference on Computer Vision*, pp. 609-623, 2012.

[11] D. Li, X. Chen and K. Huang, "Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios," *IEEE Asian Conference on Pattern Recognition*, pp. 111-115, 2015.

[12] K. Xu, J. Ba and R. Kiros, "Show, attend and tell: neural image caption generation with visual attention," *International Conference on Machine Learning*, vol. 14, pp. 77-81, 2015.

[13] O. Vinyals, A. Toshev and S. Bengio, "Show and tell: a neural image caption generator," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156-3164, 2015.

[14] J. Mao, W. Xu and Y. Yang, "Deep captioning with multimodal recurrent neural networks (m-RNN)," *International Conference on Learning Representations*, 2015.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations*, 2015.

[16] K. Cho, B. Van and C. Gulcehre, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *Empirical Methods in Natural Language Processing,* 2014.

[17] D. Bahdanau, K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *International Conference on Learning Representations*, 2015.

[18] M. Hirzer, C. Beleznai, P. M. Roth and H. Bischof, "Person re-identification by descriptive and discriminative classification," *Scandinavian Conference on Image Analysis*, pp.91-102, 2011.

[19] J. Shao, Loy C. Change and X. Wang, "Scene-independent group profiling in crowd," *IEEE Conference on Computer Vision and Pattern Recognition*, pp.2219-2226, 2014.

[20] D. Gray, S. Brennan and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," *IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, 2007.