Pedestrian Attribute Recognition: A Survey

Xiao Wang[1,2], Shaofei Zheng[1], Rui Yang[1],
Aihua Zheng[1], Zhe Chen[3], Jin Tang[1], Bin Luo[1]

1. School of Computer Science and Technology, Anhui University, Hefei, Anhui Province, China.
2. Peng Cheng Laboratory, Shenzhen, China
3. School of Computer Science, Faculty of Engineering, The University of Sydney, Australia

Abstract

Pedestrian Attribute Recognition (PAR) is an important task in computer vision community and plays an important role in practical video surveillance. The goal of this paper is to review existing works using traditional methods or based on deep learning networks. Firstly, we introduce the background of pedestrian attribute recognition, including the fundamental concepts and formulation of pedestrian attributes and corresponding challenges. Secondly, we analyze popular solutions for this task from eight perspectives. Thirdly, we discuss the specific attribute recognition, then, give a comparison between deep learning and traditional algorithm based PAR methods. After that, we show the connections between PAR and other computer vision tasks. Fourthly, we introduce the benchmark datasets, evaluation metrics in this community, and give a brief performance comparison. Finally, we summarize this paper and give several possible research directions for PAR. The project page of this paper can be found at: https://sites.google.com/view/ahu-pedestrianattributes/.

Keywords: Pedestrian Attribute Recognition; Multi-label Learning; Multi-task Learning; Deep Learning; CNN-RNN

1. Introduction

Pedestrian attributes, are humanly searchable semantic descriptions and can be used as soft-biometrics in visual surveillance, with applications in person re-identification, face verification and human identification. Pedestrian attribute recognition (PAR) aims at mining the attributes of target person whose image is given. Different from low-level features, such as HOG, LBP or deep features, attributes can be viewed as high-level semantic information which is more robust to viewpoint changes and viewing condition variations. Hence, many tasks in computer vision integrate the attribute information into their algorithms to achieve better performance, such as pedestrian detection (Chen et al., 2021), person reidentification, action recognition and scene understanding. Although many works have been proposed on this topic, however, PAR is still an unsolved problem due to challenging factors, such as view point change, low illumination, low resolution.

Traditional pedestrian attribute recognition methods usually focus on developing robust feature representation from the perspectives of hand-crafted features, powerful classifiers or attributes relations. Some milestones including HOG, SIFT, SVM or CRF model. However, the reports on large-scale benchmark evaluations suggest that the performance of these traditional algorithms is far from the requirement of realistic applications. Over the past several years, deep learning has achieved an impressive performance due to its success on automatic feature extraction using multi-layer nonlinear transformation, especially in computer vision, speech recognition and natural language processing. Many deep learning based pedestrian attribute recognition algorithms have been proposed based on these breakthroughs.

Although so many algorithms have been proposed, until now, there exists no work to make a detailed survey, comprehensive evaluation and insightful analysis on these attribute recognition algorithms. In this paper, we summarize existing works on pedestrian attribute recognition, including traditional methods and popular deep learning based algorithms, to better understand this direction and help other researchers to quickly capture main pipeline as well as latest research frontier. Specifically speaking, we attempt to address the following several important issues: 1) What is the connection and difference between traditional and deep learning-based pedestrian attribute recognition algorithms? We analyse traditional and deep learning based algorithms from different classification rules, such as part-based, group-based or end-to-end learning; 2) How the pedestrian attributes contribute to other related computer vision tasks? We also review some person attributes guided computer vision tasks, such as person re-identification, human detection, to fully demonstrate the effectiveness and widely applications in many related tasks; 3) How to make better use of deep networks for pedestrian attribute recognition and what is the future direction of the development on attribute recognition? By analysing existing person attribute recognition algorithms and some top-ranked baseline methods, we draw some useful conclusions and provide some possible research directions.

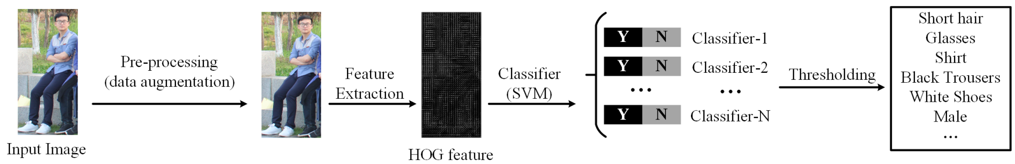

Figure 1: The regular pipeline of pedestrian attribute recognition.

2. Problem Formulation and Challenging Factors

Given a person image $\mathcal{I}$, pedestrian attribute recognition aims at predicting a group of attributes a_i to describe the characteristic of this person from a pre-defined attribute list $\mathcal{A} = \{a_1, a_2, ..., a_L\}$. This task can be handled in different ways, such as multi-label classification and binary classification. As shown in Figure 1, the input images are usually processed with data augmentation to attain more training samples. Then, the features of processed images are extracted with deep learning methods or manual designed algorithms like HOG. With the feature representation and its labels, we can train the machine learning model, such as a classifier, for each attribute in a supervised way. In the testing phase, we can use this model to predict the response score of each attribute and assume that this input image has a corresponding attribute if its score is larger than the given threshold. In addition to such simultaneous attribute prediction, there are also algorithms that predict the attribute in a recurrent way, i.e., the attributes are predicted one after another.

Although good performance has been achieved based on deep learning models, however, this task is still challenging due to the large intra-class variations in attribute categories (appearance diversity and appearance ambiguity (Deng et al., 2014)). We list challenging factors which may obviously influence the final recognize performance as follows: **1). Multi-views.** The images taken from different angles by the camera lead to the viewpoint issues for many computer vision tasks. Due to the body of human is not rigid, which further making the person attribute recognition more complicated. **2). Occlusion.** Partial occlusion of human body by other person or things increases the difficulty of person attributes recognition. Because the pixel values introduced by the occluded parts may make the model confused and lead to wrong predictions. **3). Unbalanced attribute distribution.** Each person have different attributes, therefore, the number of attributes are variable which leads to unbalanced data distribution. **4). Low resolution.** In practical scenarios, the resolution of images are rather low due to the high-quality cameras are rather expensive. **5). Illumination.** The images may taken from any time in 24 hours. Hence, the light condition is variable at different time. The shadow may also be taken in the person images and the images taken from night time maybe totally ineffective. **6). Blur.** When person is moving, the images taken by the camera may blur. Recognizing attributes in this situation is also a very challenging task.

3. The Review of PAR Algorithms

In this section, we will review existing pedestrian attribute recognition algorithms from following eight aspects: global based, local parts based, visual attention based, sequential prediction based, newly designed loss function based, curriculum learning based, graphic model based and others algorithms. A brief summary of these methods can be found in Table 2 and 3.

3.1. Global Image-based Models

(Sudowe et al., 2015) proposes multi-branch classification layers for each attribute learning with convolutional network. They adopt a pre-trained AlexNet as basic feature extraction sub-network, and replace the last fully connected layer with one loss per attribute using the KL-loss (Kullback-Leibler divergence based loss function). (Li et al., 2015) introduce deep neural network for PAR and attempt to handle the following two issues existed in traditional methods: 1). hand crafted features; 2). ignored correlations between attributes. Two algorithms DeepSAR and DeepMAR are proposed in this paper. DeepSAR do not model the correlations between human attributes which maybe the key to further improving the overall recognition performance. Therefore, they propose the DeepMAR which takes human image and its attribute label vectors simultaneously and jointly considers all the attributes via sigmoid cross entropy loss. In addition, they also consider the unbalanced label distribution in practical surveillance scenarios and propose an improved loss function which widely used in many subsequent deep PAR works. (Abdulnabi et al., 2015) propose a joint multi-task learning algorithm for attribute estimation using CNN, named MTCNN. The MTCNN lets the CNN models share visual knowledge among different attribute categories. They adopt multi-task learning on the CNN features to estimate corresponding attributes and use decomposition method to obtain shareable latent task matrix and combination matrix from total classifier weights matrix. Thus, they can achieve flexible global sharing and competition between groups through learning localized features. The Accelerate Proximal Gradient Descent algorithm is used for the optimization.

Many works adopt CNN-RNN framework to take advantage of the intra-group mutual exclusion and inter-group correlation, but they ignore the prior knowledge underlying the attribute dataset. (Kai Han, 2019) propose to explore the correlation between different attributes by mining the attribute co-occurrence prior. Specifically, they integrate the information from different predictions with an attribute aware pooling method. Their model follows multi-branch architecture and context information is gathered to improve the final recognition performance.

Summary: According to aforementioned algorithms, we can find that these algorithms all take the whole images as input and conduct multi-task learning for PAR. They all attempt to learn more robust feature representations using feature sharing, end-to-end training or multi-task learning. The benefits of these models are simple, intuitive and highly efficient which are very important for practical applications. However, the performance of these models is still limited due to the lack of consideration of fine-grained recognition.

3.2. Part-based Models

As is known to all, we can train attribute classifiers simpler if we could isolate image patches corresponding to the same body part from the same viewpoint. However, direct use object detectors is not reliable for body parts localization before the year of 2011 due to its limited ability. (Bourdev et al., 2011) adopt the *poselets* to decompose the image into a set of parts, each capturing a salient pattern corresponding to a given viewpoint and local pose. This provides a robust distributed representation of a person from which attributes can be inferred without explicitly localizing different body parts. Specifically, they first detect the poselets on given image and obtain their feature representations. Then, they train multiple SVM classifiers which are used for *poselet-level, person-level, context-level* attribute classification, respectively.

RAD* (ICCV-2013, (Joo et al., 2013)) proposes a part learning algorithm from the perspective of appearance variance while previous works focus on handling geometric variation which require manual part annotation, such as poselet (Bourdev et al., 2011). They first divide the image lattice into a number of overlapping sub-regions (named *window*). A grid of size $W \times H$ is defined and any rectangle on the grid containing one or more number of cells of the grid forms a window. The proposed method is more flexible in shape, size and location of part window while previous works (such as spatial pyramid matching structure, SPM (Lazebnik et al., 2006)) recursively divide the region into four quadrants and make all subregions are squares that do not overlap with each other at the same level.

With all these windows, they learn a set of part detectors that are spatially associated with that particular window. For each window, all corresponding image patches are cropped from training images and represented by HOG and color histogram feature descriptors. Then, K-means clustering is conducted based on the extracted features. Each obtained cluster denotes a specific appearance type of a part. They also train a local part detector for each cluster by logistic regression as a initial detector and iteratively refine it by applying it in the entire set again and updating the best location and scale to handle the issue of noisy clusters. After learning the parts at multi-scale overlapping windows, they follow the method for attribute classification proposed in the Poselet-based approach (Bourdev et al., 2011). Specifically, they aggregate the scores from these local classifiers with the weights given by part detection scores for final prediction.

PANDA (CVPR-2014, (Zhang et al., 2014)) find the signal associated with some attributes is subtle and the image is dominated by the effects of pose and viewpoint. For the attribute of *wear glasses*, the signal is weak at the scale of the full person and the appearance varies significantly with the head pose, frame design and occlusion by the hair. They think the key to accurately predicting the underlying attributes lies on locating object parts and establishing their correspondences with model parts. They propose to jointly use global image and local patches for person attributes recognition. They first detect the poselets, then adopt CNN to extract the feature representations of the local patches and whole human image. They directly feed the combined local and global features into the linear classifier which is a SVM (Support Vector Machine) for multiple attributes estimation.

AAWP (ICCV-2015, (Gkioxari et al., 2015)) is introduced to validate whether parts could bring improvements on both action and attribute recognition. The CNN features are computed on a set of bounding boxes which associated with the instance to classify, i.e., the whole instance, the oracle or person detector provided and poselet-like part detector provided. For the part detector module, they design their network by following the object detection algorithm RCNN (Girshick et al., 2015). Given the image and detected parts, they use CNN to obtain fc7 features and concatenate them into one feature vector as its final representation. Therefore, the action or attribute category can be estimated with pre-trained linear SVM classifier. This work further expanding and validating the effectiveness and necessity of parts in a more wider way.

MLCNN (ICB-2015, (Zhu et al., 2015)) propose a multi-label convolutional neural network to predict multiple attributes together in a unified framework. They divide the whole image into 15 overlapping patches and use a convolutional network to extract its deep features. They adopt corresponding local parts for specific attribute classification. They also use the predicted attributes to assist person re-identification and their experiments validate the important role of attributes in human related tasks.

ALM (ICCV-2019, (Tang et al., 2019)) predict attributes in a hierarchical manner and fuse these results with a simple voting scheme. More importantly, they propose a weakly-supervised attribute localization module (ALM) based on spatial transformer network for each branch. The ALM also contains a tiny channel-attention module for feature augmentation. Their PAR network is trained with deep supervision mechanism.

ARAP (BMVC2016, (Luwei Yang and Tan, 2016)) adopts an end-to-end learning framework for joint part localization and multi-label classification for person attribute recognition. It mainly contains the initial convolutional feature extraction layers, a key point localization network, an adaptive bounding box generator for each part, and the final attribute classification network for each part. Their network contains three loss functions, i.e., the regression loss, aspect ratio loss and classification loss. Specifically, they first extract the feature map of input image, then conduct key points localization. Given the key points, they divide human body into three main regions (including hard, torso and legs) and obtain an initial part bounding box. On the other hand, they also take previous fc7 layer's features as input and estimate the bounding box adjustment parameters. Given these bounding box, they adopt bilinear sampler to extract corresponding local features. Then, the features are fed into two fc layers for multi-label classification.

DeepCAMP (CVPR-2016, (Diba et al., 2016)) propose a novel CNN that mines mid-level image patches for fine-grained human attributes recognition. Specifically, they train a CNN to learn discriminative patch groups, named *DeepPattern*, then, utilize regular contextual information and also deploy an iteration of feature learning and patch clustering to purify the set of dedicated patches. The main insight of this paper lies on that a better embedding can help improve the quality of clustering algorithm in pattern mining algorithm. Therefore, they propose an iteration algorithm where in each iteration, they train a new CNN to classify cluster labels obtained in previous iteration to help improve the embedding. On the other hand, they also concatenate features from both local patch and global human bounding box to improve the clusters of mid-level elements.

PGDM (ICME-2018, (Li et al., 2018)) is the first work which attempts to explore the structure knowledge of pedestrian body (i.e., pedestrian pose) for person attributes learning. They first estimate the key points of given human image using pre-trained pose estimation model. Then, they extract the part regions according to these key points. The deep features of part regions and whole image are all extracted and used for attribute recognition independently. These two scores are then fused together to achieve final attribute recognition. The attribute recognition algorithm contains two main modules: i.e., the main net (AlexNet) and PGDM. The introduced PGDM module is an existing pose estimation algorithm. They directly train a regression network to predict the pedestrian pose with coarse ground truth pose information which obtained from existing pose estimation model. Then, they transform the key points into informative regions using spatial transformer network, and use independent neural network for feature learning from each key point related region. They jointly optimize the main net, PGDM and pose regression network.

DHC (ECCV-2016, (Li et al., 2016)) propose to use *deep hierarchical contexts* to help person attribute recognition due to the background would sometimes provide more information than target object only. Specifically, the *human-centric context* and *scene context* are introduced in their network architecture. They first construct input image pyramid and pass them all through VGG-16 to obtain multi-scale feature maps. They extract features of four set of bounding box regions, i.e., the whole person, detected parts of target object, nearest neighbour parts from the image pyramid and global image scene. The first two branches (the whole person and parts) are regular pipeline for person attributes recognition algorithm. The main contributions of this paper lie on the later two branches, i.e., the human-centric and scene-level contexts help improve the recognition results. Once the scores of these four branches are obtained, they sum up all the scores as final attribute score. Due to the use of context information, this neural network needs more external training data than regular pedestrian attribute recognition task. For example, they need to detect the part of human body (head, upper and bottom body regions) and recognize the style/scene of given image. They propose a new dataset named *WIDER*, to better validate their ideas. Although the human attribute recognition results can be improved significantly via this pipeline, however, this model looks a little more complicated than other algorithms.

LGNet (BMVC-2018, (Liu et al., 2018)) propose a Localization Guide Network (LGNet) which can localize the areas corresponding to different attributes. It also follows the local-global framework. Specifically, they adopt Inception-v2 as their basic CNN model for feature extraction. For global branch, they adopt global average pooling layer (GAP) to obtain its global features. Then, a fc layer is utilized to output its attribute predictions. For the local branch, they use 1×1 convolution layer to produce c class activation maps for each image. Then, they capture an activation box for each attribute by cropping the high-response areas of the corresponding activation map. They also use EdgeBoxes to generate region proposals to obtain local features from the input image. In addition, they also consider the different contributions of extracted proposals and different attributes should focus on different local features. Therefore, they use the class active map for each attribute to serve as a guide to determine the importance of the local features to different attributes. Finally, the global and attended local features are fused together by element-wise sum for PAR.

Summary: Based on the reviewed papers in this subsection, it is intuitive to find that these algorithms all adopt both global and fine-grained local features. The localization of body parts is achieved via an external part localization module, such as part detection, pose estimation, poselets or proposal generation algorithm. The use of part information improves the overall recognition performance significantly. At the same time, it also brings some shortcomings as follows: Firstly, as an operation in the middle phase, the final recognition performance heavily relies on the accuracy of part localization. In another word, the inaccurate part detection results will bring the wrong features for final classification. Secondly, it also needs more training or inference time due to the introducing of human body parts. Thirdly, some algorithms need manual annotated labels for part location which further increasing the cost of manpower and money.

3.3. Attention-based Models

HydraPlus-Net (ICCV-2017, (Liu et al., 2017)) is introduced to encode multi-scale features from multiple levels for pedestrian analysis using multi-directional attention (MDA) modules. It contains two main modules, i.e., the Main Net (M-net) which is a regular CNN and the Attentive Feature Net (AF-net) which includes multiple branches of multi-directional attention modules applied to different semantic feature levels. The AF-net and M-net share same basic convolution architectures and their outputs are concatenated and fused by global average pooling and fc layers. The output layer can be the attribute logits for attribute recognition or feature vectors for person re-identification. In another word, it can be used to minimize the cross-entropy loss and softmax loss for PAR and person re-identification respectively.

VeSPA (arXiv-2017, (Sarfraz et al., 2017)) takes the view cues into consideration to better estimate corresponding attribute. Because the authors find that the visual cues hinting at attributes can be strongly localized. Besides, the inference of person attributes such as hair, backpack, shorts, are highly dependent on the acquired view of the pedestrian. The image is fed into the Inceptions networks and its feature representation can be obtained. The view-specific unit is introduced to mapping the feature maps into coarse attribute prediction. Then, a view predictor is used to estimate the view weights. The attention weights are used to multiply view-specific predictions and obtain the final multi-class attribute prediction. The view classifier and attribute predictors are trained with separate loss function. The whole network is an unified framework and can be trained in an end-to-end manner.

DIAA (ECCV-2018, (Sarafianos et al., 2018)) can be seen as an ensemble method for person attribute recognition. Their model contains a multi-scale visual attention and a weighted focal loss for deep imbalanced classification. For the multi-scale visual attention, the authors adopt feature maps from different layers. They propose the weighted focal loss function to measure the difference between predicted attribute vectors and ground truth. In addition, they also propose to learn the attention maps in a weakly supervised manner (only the attribute labels, no specific bounding box annotation) to improve the classification performance by guiding the network to focus its resources to those spatial parts that contain information relevant to the input image. The attention sub-network takes the feature map as input and output an attention mask. The output is then fed to attention classifier to estimate the pedestrian attributes.

CAM (PRL-2017, (Guo et al., 2017)) propose to use and refine attention map to improve the performance of PAR.Their model contains two main modules, i.e., the multi-label classification sub-network and attention map refinement module. The adopted CAM net also follows the category-specific framework, in another word, different attribute classifiers have different parameters for the fc layer. They use the parameters in fc layer as weights to linearly combine the feature maps from the last convolutional layer to get the attention of each category. However, this naive implementation of attention mechanism could not focus on the right regions all the time due to low resolution, over-fitting training. To handle this issue, they exploring refine the attention map by tuning CAM network. They measure the appropriateness of an attention map based on its concentration and attempt to make the attention map to highlight a smaller but concentrated region. Specifically, they introduce a weighted average layer to obtain attention map first. Then, they use average pooling to down-sample its resolution to capture the importance of all the potential relevant regions. After that, they also adopt softmax layer to transform the attention map into a probability map. Finally, the maximum probability can be obtained via the global average pooling layer. On the basis of the maximum probability, the authors propose the *exponential loss function* to measure the appropriateness of the attention heat map. For the training of the network, the authors first pre-training the CAM network only by minimizing classification loss; then, they adopt joint loss functions to fine-tuning the whole network.

JLPLS-PAA (TIP-2019, (Tan et al., 2019)) explore multiple attention mechanisms to select important and discriminative regions or pixels to handle the issues such as large pose variations, clutter background. Different from regular spatial, temporal or channel-view, they propose the parsing attention, label attention and spatial attention. Specifically, the parsing model is used to locate the specific body regions at pixel-level in a split-and-aggregate way. The label attention is formulated by assigning several attention maps for each label under image-level supervisions. The spatial attention is also considered to locate the most discriminative image regions for all attributes with image-level supervisions. It is worthy to note that this work is the first attempt to jointly learn multiple attention mechanisms in a multi-task-like learning manner.

IA^2**-Net (PRL-2019, (Ji et al., 2019))** propose an image-attribute reciprocal guidance representation (RGR) method to investigate image-guided feature and attribute-guided feature. Their method is developed based on the following observation: some attributes are concrete, such as "Hair Style, Shoes Style", but some are abstract attributes (For example, "Age Range, Role Types"). They also develop a fusion attention mechanism to assign different attentions to different RGR features. Besides, they combine the focal loss and cross-entropy loss to handle the attribute imbalance problem.

Da-HAR (AAAI-2020, (Wu et al., 2019)) attempt to recognize the human attributes based on coarse-to-fine framework with self-mask operator. Their self-mask block is trained on MS-COCO dataset and used for person segmentation. With the help of a mask, their model is insensitive to distraction and clutter background. Hierarchical features from various layers of backbone network are fused with 1×1 operator and attention module. The predictions from such side branch are fused with the main branch for final decision making.

CAS (ICME-2020, (Zeng et al., 2020)) A Co-Attentive Sharing module is proposed by (Zeng et al., 2020) based on soft-sharing structure in multi-task learning, which could mine discriminative channels and spatial regions for more effective feature sharing. More detail, synergistic branch, attentive branch and task-specific branch are explored for each layer, then, the results of three branches are aggregated as the input features for the subsequent layer of each task.

(Zhang et al., 2019) propose the task-aware attention mechanism (named TAN) to explore the importance of each position across different tasks. They first use a cloth detector to crop out the target region, then, extract its feature with CNN.

The spatial attention and task attention modules are employed to learn feature maps and the t-distribution Stochastic Triplet Embedding (t-STE) loss function is used for the optimization.

Summary: Visual attention is a hot research topic in current deep learning era and has been widely used in many domains. Generally speaking, attention is the behavioral and cognitive process of selectively concentrating on a discrete aspect of information, whether deemed subjective or objective, while ignoring other perceivable information [1]. Pedestrian attribute recognition also follows this framework and aforementioned works also validate the effectiveness of attention mechanism. However, the works integrate with attention mechanism are still limited. How to design new attention models or directly borrow existing attention algorithms from other domains is still unexplored.

3.4. Sequential Prediction based Models

CNN-RNN (CVPR-2016, (Wang et al., 2016)) Regular multi-label image classification framework learn independent classifier for each category and employ ranking or threshold on the classification results, fail to explicitly exploit the label dependencies in an image. This paper first adopts RNNs to address this problem and combine with CNNs to learn a joint image-label embedding to characterize the semantic label dependency as well as the image-label relevance. This model can model the label co-occurrence dependencies in the joint embedding space by sequentially linking the label embeddings. For the inference of CNN-RNN model, they attempt to find the sequence of labels that maximize the prior probability. The training of the CNN-RNN model can be achieved by cross-entropy loss function and back-propagation through time (BPTT) algorithm.

JRL (ICCV-2017, (Wang et al., 2017)) firstly analyse existing learning issues in the pedestrian attribute recognition task, e.g., poor image quality, appearance variation and little annotated data, and propose to explore the interdependency and correlation among attributes and visual context as extra information source to assist attribute recognition. Hence, the JRL model is proposed to joint recurrent learning of attribute context and correlation, as its name shows. To better mine these extra information for accurate person attribute recognition, the authors adopt *sequence-to-sequence* model to handle aforementioned issues. They first divide the given person image into multiple horizontal strip regions and form a region sequences in top-bottom order. The obtained region sequences can be seen

[1] https://en.wikipedia.org/wiki/Attention

as the input sentence in natural language processing, and can be encoded with the LSTM network in a sequential manner. In decoding phase, the decoder LSTM takes both *intra-person attribute context* and *inter-person similarity context* as input and output variable-length attributes over time steps. The attribute prediction in this paper can also be seen as a generation scheme. To better focus on local regions of person image for specific attributes and obtain more accurate representation, they also introduce the attention mechanism to attend the intra-person attribute context.

GRL (IJCAI-2018, (Zhao et al., 2018)) is developed based on JRL which also adopts the RNN model to predict the human attributes in a sequential manner. Different from JRL, GRL is formulated to recognize human attributes by group, and gradually pay attention to both intra-group and inter-group relationships. They divide the whole attribute list into many groups because the attributes in intra-group are mutual exclusive and also correlated between inter-group. For example, *BoldHair* and *BlackHair* cannot occur on the same person image, but they are both related to the head-shoulder region of a person and can be in the same group to be recognized together. It is an end-to-end single model algorithm with no need for preprocessing and it also exploits more latent intra-group and inter-group dependency among grouped pedestrian attributes.

JCM (arXiv-2018, (Liu et al., 2018)) Existing sequential prediction based PAR algorithms, such as JRL, GRL, may be easily influenced by different manual division and attributes orders due to the weak alignment ability of RNN. This paper proposes a joint CTC-Attention model (JCM) to conduct attribute recognition, which could predicts multiple attribute values with arbitrary length at a time avoiding the influence of attribute order in the mapping table.

JCM is actually a multi-task network which contains two tasks: the attribute recognition and person re-identification. They use ResNet-50 as the basic model to extract features for both tasks. For the attribute recognition, they adopt the Transformer as their attention model for the alignment of long attribute sequence. And the connectionist temporal classification (CTC) loss and cross entropy loss functions are used for the training of network. For the person re-ID, they directly use two fully connected layers to obtain feature vectors and use softmax loss function to optimize this branch. In the test phase, the JCM could simultaneously predicts the person identity and a set of attributes. They also use beam search for the decoding of attribute sequence. Meanwhile, they extracts the features from the CNN in base model to classify pedestrians for person re-ID task.

RCRA (AAAI-2019, (Xin Zhao and Yan, 2019)) propose two models, i.e., Recurrent Convolutional (RC) and Recurrent Attention (RA) for pedestrian attribute recognition. The RC model is used to explore the correlations between different attribute groups with Convolutional-LSTM model and the RA model takes the advantage of the intra-group spatial locality and inter-group attention correlation to improve the final performance. Specifically, they first divide all the attributes into multiple attribute groups, similar with GRL. For each pedestrian image, they use CNN to extract its feature map and feed it to ConvLSTM layer group by group. Then, new feature map for each time step can be obtained by adding a convolutional network after ConvLSTM. Finally, the features are used for attribute classification on current attribute group. Based on aforementioned RC model, they also introduce visual attention module to highlight the region of interest on the feature map. The attended feature maps are used for final classification. The training of this network is also based on weighted cross-entropy loss function proposed in WPAL-network.

Summary: As we can see from this subsection, these algorithms all adopt the sequential estimation procedure. Because the attributes are correlated to each other, and they also have various difficulties. Therefore, it is an interesting and intuitive idea to adopt the RNN model to estimate the attributes one by one. Among these algorithms, they integrate different neural networks, attribute groups, multi-task learning into this framework. Compared with CNN based methods, these algorithms are more elegant and effective. The disadvantage of these algorithms is the time efficiency due to the successive attribute estimation. In the future works, more efficient algorithms for the sequential attributes estimation are needed.

3.5. Newly Designed Loss Function based Models

WPAL-network (BMVC-2017, (Zhou et al., 2017)) is proposed to simultaneously recognize and locate the person attributes in a weakly-supervised manner (i.e., only person attribute labels, no specific bounding box annotation). The GoogLeNet is adopted as their basic network for feature extraction. They fuse features from different layers and feed them into Flexible Spatial Pyramid Pooling layer (FSPP). The outputs of each FSPP are fed into fully connected layers and output a vector whose dimension is same as the number of pedestrian attributes. In addition, the authors also introduce a novel weighted cross entropy loss function to handle the extremely imbalanced distribution of positive and negative samples of most attribute categories.

AWMT (MM-2017, (He et al., 2017)) As is known to all, the learning difficulty of various attributes is different. However, most of existing algorithms ignore this situation and share relevant information in their multi-task learning framework. This will leads to *negative transfer*, in another word, the inadequate brute-force transfer may hurt the learner's performance when two tasks are dissimilar. AWMT proposes to investigate a shared mechanism that is possible of *dynamically* and *adaptively* coordinating the relationships of learning different person attribute tasks. Specifically, they propose an adaptively weighted multi-task deep framework to jointly learn multiple person attributes, and a validation loss trend algorithm to automatically update the weights of weighted loss layer.

They use ResNet-50 as backbone network and take both train and val images as input. The basic network will output its predicted attribute vectors for both train and val images. Hence, the train loss and val loss can be obtained simultaneously. The val loss is used to update the weight vectors which are then utilized to weight different attributes learning. They propose the validate loss trend algorithm to adaptively tuning the weight vector. The intuition behind their algorithm is, when learning multiple tasks simultaneously, the "important" tasks should be given higher weights to increase the scale of loss of the corresponding tasks.

ArXiv-2019, (Yaghoubi et al., 2020) is the first work which utilize the *hard* attention to address the influence of background using binary mask predicted by mask R-CNN. Then, they train their network based on the multi-task learning to capture the semantic dependencies between most of the labels. The authors define a weighted sum loss function to consider various contributions of each category in the loss value.

HFE (CVPR-2020, (Yang et al., 2020)) introduces external person ID constraints for hierarchical feature embedding (HFE) based on newly designed HFE loss. This loss function is extended from triplet loss function and consists of inter-triplet loss, intra-triplet loss and absolute boundary regularization. Therefore, each class could gather more compactly, leading to a more distinct boundary between classes.

Meanwhile, (Jia et al., 2020) argue that existing setting of PAR is not practical because of the large number of identical pedestrian identities in train and test set. They re-divide the dataset to ensure that the images with the same person ID do not occur in train and test set simultaneously, and implement a strong baseline method based on this setting. Their experimental results demonstrate that existing PAR algorithms are overclaimed. They think distinguish the fine-grained attributes in the same area (such as sandals *vs.* sneakers) is more important than locating the area of the specific attribute.

(Ji et al., 2020) propose the **MTA-Net** to address complex relations between images and attributes, and imbalanced distribution of pedestrian attributes. They jointly use the knowledge of previous, current and next time steps based on CNN-RNN framework. Besides, the focal balance loss (FBL) function is proposed to handle the second issue.

Summary: There are few works focus on designing new loss functions for pedestrian attribute recognition. WPAL-network (Zhou et al., 2017) consider the unbalanced distribution of data and propose a weighted cross-entropy loss function according to the proportion of positive labels over all attribute categories in the training dataset. This method seems a little tricky but has been widely used in many PAR algorithms. AWMT (He et al., 2017) propose an adaptive weighting mechanism for each attribute learning to make the network focus more on handling the "hard" tasks. These works full demonstrate the necessity of designing novel loss functions to better train the PAR network.

3.6. Curriculum Learning based Algorithms

MTCT (WACV-2017, (Dong et al., 2017)) proposes a multi-task curriculum transfer network to handle the issue on the lack of manually labelled training data. Their algorithm contains multi-task network and curriculum transfer learning. For the multi-task network, they adopt five stacked Network-In-Network (NIN) convolutional units and N parallel branches, with each branch representing a three layers of fully connected sub-network for modelling one of the N attributes respectively. Softmax loss function is adopted for the model training.

Cognitive studies suggest that a better learning strategy adopted by human/animals is to start with learning easier tasks before gradually increasing the difficulties of the tasks, rather than blindly learn randomly organised tasks. Therefore, they adopt curriculum transfer learning strategy for clothing attribute modelling. Specifically, it is consisted of two main stages. In the first stage, they use the clean (easier) source images and their attribute labels to train the model. In the second stage, they embed cross-domain image pair information and simultaneously append harder target images into the model training process to capture harder cross-domain knowledge. They adopt t-STE (t-distribution stochastic triplet embedding) loss function to train the network

CILICIA (ICCV-2017, (Sarafianos et al., 2017)) Similar with MTCT (Dong et al., 2017), CILICIA also introduces the idea of curriculum learning into person attribute recognition task to learn the attributes from easy to hard. They explore the correlations between different attribute learning tasks and divide such correlations into strongly and weakly correlated tasks. Specifically, under the framework of multi-task learning, they use the respective Pearson correlation coefficients to measure the strongly correlated tasks. For the multi-task network, they adopt the categorical cross-entropy function (Zhu et al., 2017) to measure the difference between predictions and targets. To weight different attribute learning tasks, one intuitive idea is to learn another branch network for weights learning. They adopt the *supervision transfer* learning technique to help attribute learning in weakly correlated group.

They also propose CILICIA-v2 (Sarafianos et al., 2018) by introducing an effective method to obtain the groups of tasks using hierarchical agglomerative clustering. It can be any number and not just only two groups (i.e., strong/weakly correlated).

DCL (ICCV-2019, (Wang et al., 2019)) introduces an unified framework, named dynamic curriculum learning, to online adaptively adjust the sampling strategy and loss learning in a batch to handle the issues caused by imbalanced data distribution. Specifically, they design two level curriculum schedulers: sampling scheduler and loss scheduler. The first one aims at finding the most meaningful samples in one batch to learn from imbalanced to balanced distribution and easy to hard. The second one is used to achieve a good trade-off between classification and metric learning loss. They achieve new state-of-the-art recognition performance on two attribute datasets.

Summary: Inspired by recent progress of cognitive science, the researchers also consider using such "easy" to "hard" learning mechanism for PAR. They introduce existing curriculum learning algorithm into their learning procedure to model the relations between each attribute. This makes the PAR algorithms look more intelligent due to the ability of estimating the "easier" attributes first just like humans. Some other algorithms such as self-paced learning are also used to model the multi-label classification problem or other computer vision tasks. It is also worthy to introduce more advanced works of cognitive science to guide the learning of PAR. In addition, the meta-learning has shown its ability to "learning to learn" in many tasks, such as fine-grained classification, few-shot learning. It will also be an interesting research direction to integrate this learning framework for PAR.

3.7. Graphic Model based Algorithms

Graphic models are commonly used to model structure learning in many applications. Similarly, there are also some works to integrate these models into the PAR task.

DCSA* (ECCV-2012, (Chen et al., 2012)) propose to model the correlations between human attributes using conditional random field (CRF). They first estimate the pose information and locate the local parts of upper body only. Then, four types of base features are extracted from these regions. These features are fused to train multiple attribute classifiers via SVM. The key idea of this paper is to apply the fully connected CRF to explore the mutual dependencies between attributes. They treat each attribute function as a node of CRF and the edge connecting every two attribute nodes reflects the joint probability of these two attributes. The belief propagation is adopted to optimize the attribute label cost.

A-AOG* (TPAMI-2018, (Park et al., 2018)) is short for attribute And-Or grammar, which is proposed explicitly to represent the decomposition and articulation of body parts, and account for the correlations between poses and attributes. This algorithm is developed based on And-Or graph and the and-nodes denote decomposition or dependency; the or-nodes represent alternative choices of decomposition or types of parts. Specifically speaking, it mainly integrates the three types of grammars: *phrase structure grammar, dependency grammar* and an *attribute grammar*. They use deep CNN to generate proposals for each part and adopt greedy algorithm based on the beam search to optimize aforementioned objective function.

VSGR (AAAI-2019, (HUANG, 2019)) propose to estimate the pedestrian attributes via visual-semantic graph reasoning (VSGR). They argue that the accuracy of person attribute recognition is heavily influenced by: 1). only local parts are related with some attributes; 2). challenging factors, such as pose variation, viewpoint and occlusion; 3). the complex relations between attributes and different part regions. Therefore, they propose to jointly model spatial and semantic relations of region-region, attribute-attribute, and region-attribute with a graph-based reasoning framework.

This algorithm mainly contains two sub-networks, i.e., the visual-to-semantic sub-network and semantic-to-visual sub-network. For the first module, it first divides the human image into a fixed number of local parts. They construct a graph whose node is the local part and edge is the similarity of different parts. Different from regular relation modelling, they adopt both the similarity relations between parts and topological structures to connect one part with its neighbour regions. The two sub-graphs are combined to compute the output of spatial graph. The semantic-to-visual sub-network can also be processed in similar manner and it also outputs sequential attribute prediction. The outputs of these two sub-networks are fused as the final prediction and can be trained in an end-to-end way.

JLAC (AAAI-2020, (Tan et al., 2020)) propose the JLAC (Joint Learning of Attribute and Contextual relations) for PAR which contains two main modules: Attribute Relation Module (ARM) and Contextual Relation Module (CRM). The ARM module is used to explore the correlations among multiple attributes based on an attribute graph with attribute-specific features. For the CRM, the authors construct a graph projection scheme that targets at project the 2-D feature map into a set of nodes from different image regions. This module fully explored the contextual relations among those regions. The GCN is adopted to mine the graph structured features for the two modules and the whole architecture can be optimized in an end-to-end manner.

BCRNNs (CVPR-2018, (Wang et al., 2018)) propose to use Bidirectional Convolutional Recurrent Neural Networks (BCRNNs) to address the problem of visual fashion analysis based on their defined grammar topologies. Specifically, their proposed dependency grammar could capture kinematics-like relations, and symmetry grammar can accounting for the bilateral symmetry of clothes.

Summary: Due to the relations existed in multiple attributes, many algorithms are proposed to discover such information for PAR. Therefore, the Graphic models are easily introduced into the learning pipeline, such as Markov Random Field, Conditional Random Field, And-Or-Graph or Graph Neural Networks. The works reviewed in this subsection are the outputs by integrating the graphic models with PAR. Maybe the other graphic models can also be used for PAR to achieve better recognition performance. Although these algorithms have so many advantages on model the relations between pedestrian attributes, however, these algorithms seem more complex than others. The efficiency issue is also needs to be considered in practical scenarios.

3.8. Other Algorithms

This subsections are used to demonstrate algorithms that not suitable for aforementioned categories, including: PatchIt (Sudowe and Leibe, 2016), FaFS (Lu et al., 2017), GAM (Fabbri et al., 2017) and IFSL (Liuyu Xiang, 2019).

PatchIt proposes a self-supervised pre-training approach, named PatchTask, to obtain weight initializations for the PAR. It's key insight is to leverage data from the same domain as the target task for pre-training and it only relies on automatically generated rather than human annotated labels.

FaFS is proposed to design compact multi-task deep learning architecture automatically. This algorithm starts with a thin multi-layer network and dynamically widens it in a greedy manner during training. This will create a tree-like deep architecture by repeating above widening procedure and similar tasks reside in the same branch until at the top layer.

GAM proposes to handle the issue of occlusion and low resolution of pedestrian attributes using deep generative models. Specifically, their overall algorithm contains three sub-networks, i.e., the attribute classification network, the reconstruction network and super-resolution network. For the attribute classification network, they also adopt joint global and local parts for final attribute estimation.

To handle the occlusion and low-resolution problem, they introduce the deep generative adversarial network (Mirza and Osindero, 2014) to generate re-constructed and super-resolution images. And use the pre-processed images as input to the multi-label classification network for attribute recognition.

(Liuyu Xiang, 2019) propose the IFSL to handle the few-shot pedestrian attribute recognition problem. Because most previous PAR algorithms are designed for a fixed set of attributes and unable to handle the incremental few-shot learning scenario. This work introduces an extra module named attribute prototype generator, which can be seen as a high-level meta-learner that extracts the multiple-attribute information from the feature embedding. And it can produce discriminative attribute prototype embedding and therefore provide the classification weights for the novel attributes.

(Zhang et al., 2020) propose the TS-FashionNet, i.e. the Texture and Shape biased Two-Stream Networks, for fashion image analysis. Specifically, the shape-biased stream contains a landmark branch to help extract shape features; while the texture-biased stream is used to emphasize on the extraction of texture features. Then, these two branches are concatenated together to predict the clothing attributes and classify the clothes categories.

(Jia et al., 2021) argue that current evaluation for PAR is not consistent with practical scenarios and advocate zero-shot pedestrian identity setting. They propose two new dataset $PETA_{ZS}$ and RAP_{ZS} for the evaluation.

4. Discussion

In this section, we will first discuss the specific attribute recognition in this section, then, we will give a comparison between deep learning and traditional algorithm based PAR methods. After that, we will show the connections between PAR and other computer vision tasks.

4.1. Specific Attribute Recognition

In addition to the attribute recognition on whole body, there are also some attribute recognition algorithms focus on local parts of people, for example, face attribute recognition (e.g., gender, age, race). In this subsection, we will give a brief review on specific attribute recognition algorithms. For a more detailed introduction for face attribute recognition, please refer to the (Zheng et al., 2018) and (Fasel and Luettin, 2003).

(Rodríguez et al., 2017) is proposed to discover the most informative and reliable parts of a given face for improving age and gender classification. Specifically, it is a feedforward attention mechanism and mainly consists of three modules: an attention CNN, a patch CNN and a multi layer perceptron (MLP). The two CNN modules are used to predict the best attention grid to perform the glimpses and evaluate the higher resolution patches based on their importance predicted by the attention grid, respectively. The MLP module is used to integrate the information from both CNNs and make the final classifications. (Li et al., 2017) propose cumulative hidden layer and comparative ranking layer to combat the sample imbalance problem and learn more effective aging features. The cumulative hidden layer is supervised by a point-wise cumulative signal which encodes the target age labels continuously. The comparative ranking layer is supervised by a pairwise comparative signal, in another word, who is older. This is inspired by the observation that it is easier to tell which one is older given two faces than tell its accurately age. (Xing et al., 2017) conduct a comprehensively diagnose on the training and evaluating procedures of deep leaning methods for age estimation. They achieve state-of-the-art performance by following previous work with appropriate problem formulation and loss function. They also consider various factors to build a better age estimation model based on multi-task learning framework, such as the strategies to incorporate information like race and gender. Their studies are helpful to get better understandings of a deep age estimation algorithm. (Antipov et al., 2017) shed light on some open questions of human demographics estimation to improve the existing CNN-based approaches for gender and age prediction. Their work analyse four important factors of the CNN training: the target age encoding and loss function, the CNN depth, the pre-training, the training strategy. Then, they deign their model based on these experiments and achieve state-of-the-art performance. (Liu et al., 2017) propose a group-aware deep feature learning approach for facial age estimation. Specifically, they split ordinal ages into a set of discrete groups and learn deep feature transformations across age groups to project each face pair into the new feature space. They simultaneously minimize the intra-group variances of positive face pairs and maximize the inter-group variances of negative face pairs. (Chen et al., 2017) propose an approach to automatically discover "spectral attributes" which avoids manual work required for defining hand-crafted attribute representations. (Fasel and Luettin, 2003) conduct an review on automatic facial expression analysis including: facial motion, deformation extraction approaches and classification methods. (Hadid and Pietikäinen, 2009) investigate the combination of facial appearance and motion for face analysis in videos. They are inspired by the psychophysical finds which state that facial movements can provide valuable information to face analysis. They design an extended set of volume local binary patterns as well as a boosting scheme for spatio-temporal face and gender recognition from videos.

There are also some works focusing on backpack detection given a human image, for example, (Branca et al., 2002), (Ghadiri et al., 2019), (Damen and Hogg, 2011), (Ghadiri et al., 2016). The regular pipeline of these methods is to detect the human body first, then segment the carried object in a fine-grained manner.

4.2. Comparison between Deep Learning and Traditional based Algorithm

Before the deep neural network based algorithms take over the PAR community, most of traditional approaches follow a standard pipeline, which can be found in Fig. 1. Usually, we need to first conduct some pre-processing to augment the dataset, such as flip, rotation, scale variation, crop, translation, add Gaussian noise. Then, manual designed features (for example, HOG or SIFT features) are extracted to represent the person image. After that, multiple classifiers are trained to discriminate all the pedestrian attributes, such as support vector machine. In the test phase, we need to set a threshold to give an estimation whether corresponding attribute exists or not.

According to aforementioned PAR algorithms including traditional methods and deep learning based approaches, we can find the following observations: 1). Both methods all attempt to handle the PAR from the fine-grained perspective, such as estimate the attributes from local human body. The major difference lies on how to locate these regions: traditional methods rely on object detector, while deep learning methods employ more advanced object detector, visual attention mechanisms or some other information obtained from auxiliary task (for example, pose estimation). 2). Both methods all need the powerful feature representation of pedestrian images. Traditional approaches use the manual designed features, while deep learning based algorithms could learn the deep features automatically from given training dataset. This is also one of the most unique characteristics of deep learning based PAR algorithms. 3). Both methods all attempt to utilize the prior information or relations between human attributes to augment the final recognition performance. Traditional methods usually adopt graphical models such as conditional random field, markov random field as post-processing, while deep learning based algorithms can integrate such relations into their pipeline and learning in an end-to-end manner based on graph neural networks.

Generally speaking, traditional and deep learning based PAR algorithms all share similar ideas, but deep learning methods always achieve better recognition accuracy than traditional algorithms. We think one of the most important and intuitive reasons is the powerful deep features which can learn from large scale datasets. Another reason is that many challenges of PAR are hard to be modelled with traditional algorithms, but this is easy to be implemented with deep neural networks. The third reason is that deep neural networks can be integrated with traditional methods, i.e., the mode of "deep + X". This will further extending the applications of deep neural networks.

4.3. Connections between PAR and Other Tasks

Visual attributes can be seen as a kind of mid-level feature representation which may provide important information for high-level human related tasks, such as person re-identification, pedestrian detection, person tracking, person retrieval, human action recognition and scene understanding.

For the pedestrian detection, regular algorithms treat it as a single binary classification task, while (Tian et al., 2015) propose to jointly optimize person detection with semantic tasks to address the confusion of positive and hard negative samples. They use existing scene segmentation dataset to transfer attribute information to learn high-level features from multiple tasks and dataset sources.

For the person re-identification, pedestrian attributes can be seen as a kind of middle-level representation and share a common target at the pedestrian description with person re-ID. PAR focuses on local information mine while person re-identification usually capture the global representations of a person. There are already many works attempting to integrate the PAR into their person re-ID system. For example, (Lin et al., 2019) propose an attribute-person recognition network, a multi-task network which learns a re-ID embedding and predicts person attributes simultaneously. (Han et al., 2018) propose an attribute-aware attention model to learn local attribute and global category representation simultaneously in an end-to-end fashion. (Su et al., 2016) also propose to integrate the mid-level attributes into person re-identification framework and train the attribute model in a semi-supervised manner. Specifically, they first pre-train the deep CNN on an independent attribute dataset, then, fine-tuned on another dataset only annotated with person IDs. After that, they estimate attribute labels for target dataset using the updated deep CNN model. (Khamis et al., 2014) propose to integrate a semantic aspect into regular appearance-based methods. They jointly learn a discriminative projection to a joint appearance-attribute subspace, which could effectively leverage the interaction between attributes and appearance for matching. (Li et al., 2015) also present a comprehensive study on clothing attributes to assist person re-ID. They first extract the body parts and their local features to alleviate the pose-misalignment issues. Then, they propose a latent SVM based person re-ID approach to model the relations between low-level part features, middle-level clothing attributes and high-level re-ID labels of person pairs. They treat the clothing attributes as real-value variables instead of using them as discrete variables to obtain better person re-ID performance. (Layne et al., 2012) and (Layne et al., 2014) are all learn an attribute-center representation to describe people and a metric to compare attribute profiles. (Layne et al., 2012) also achieve better re-ID performance by learning a selection and weighting of mid-level semantic attributes for the description of people. (Schumann and Stiefelhagen, 2017) first train an attribute classifier and take its responses into the learning of person re-ID model based on CNNs. (Li et al., 2019) find that attributes are related to specific local regions and utilize the attribute detection to generate corresponding attribute-part detectors. This will handle the body part misalignment problem significantly for the re-ID task. (Ling et al., 2019) propose a multi-task learning network with multiple classification and verification losses for person re-ID which closely combine person identity and pedestrian attribute task. In (Su et al., 2017), the authors use the idea of multi-shot re-identification for person re-ID instead of a single prob image. Specifically, they utilize low-level features, attributes and inter-attribute correlations to make their model robust under the multi-camera setting. (Chen et al., 2018) also develop a CNN-based pedestrian attribute-assisted person re-identification framework. They first learn the attribute with a part-specific CNN and fuse them with low-level robust LOMO features. Then, they merge the learned attribute CNN embedding with identification CNN embedding under a triplet structure for person re-ID.

There are also some works integrating pedestrian attributes for person retrieval and human active recognition. For the person retrieval, (Wang et al., 2013) leverage low-level features (e.g., color) and high-level features (i.e. the person attributes) of clothing to tackle the issues caused by geometric deformation, occlusion and clutter background. Their content-based image retrieval algorithm is developed based on the bag-of-visual-words model. More importantly, they propose a re-ranking approach to improve the search result by exploiting attributes, such as the type of clothing, sleeves and patterns. (Chen et al., 2015) approach the problem of describing people by first mining clothing attributes with fine-grained attribute labels from online shopping stores. Then, they use a double-path deep domain adaptation network to bridge the gap between the collected images and practical testing data. Their work validate the effectiveness and importance of person attributes for people describe. For the human active recognition, there is a literature review summarized by (Ziaeefard and Bergevin, 2015) which also mention that the attributes are an element of semantic space and are effective features describing a basic or an intrinsic characteristic of an activity. In addition, (Liu et al., 2011) validate that attributes enable the construction of more descriptive models for human action recognition. They select attributes in a discriminative fashion or coherently integrate with data-driven attributes to make the attribute set more descriptive.

Due to the pedestrian attribute recognition is mainly focus on the clothing feature studied in many other research topics, such as part-detection, pose estimation (Murphy-Chutorian and Trivedi, 2008) and human parsing (Huang et al., 2018). But these tasks have their own emphasized point, for example: part-detection aims at locating the local parts of object using a bounding box; pose estimation focuses on locating the key points of people which will be useful for human activity recognition; And human parsing is a more fine-grained pixel-wise segmentation of human body which is more difficult than pedestrian attribute recognition. However, these tasks can be learned in a joint manner due to these tasks are all focus on human body and also have their own emphasized point. Actually, the multi-task learning has been studied for a long time in machine learning, pattern recognition and computer vision community. The joint learning of pedestrian attribute recognition and other tasks also validate the effectiveness of such multi-task setting, such as joint PAR and person re-ID algorithms described above.

5. Benchmarks

5.1. Datasets

Unlike other tasks in computer vision, for pedestrian attribute recognition, the annotation of dataset contains many labels at different levels. For example, hair style, color, hat and glass, are seen as specific low-level attributes and correspond to different areas of the images; while some attributes are abstract concepts, such as gender, orientation and age, which do not correspond to certain regions, we consider these attributes as high-level attributes. Furthermore, human attribute recognition is generally severely affected by environmental or contextual factors, such as viewpoints, occlusions and body parts. In order to facilitate the study, some datasets provide annotations of perspective, parts bounding box, occlusion.

By reviewing related work in recent years, we have found and summarized several datasets which are used to research pedestrian attribute recognition. As shown in Table 1, we only show some important parameters of these benchmark datasets, such as image numbers, attribute numbers, image source and corresponding project pages due to the limited space in this paper. For more detailed information of these datasets, please visit our project page for the arXiv version (Xiao et al., 2019).

Table 1: An overview of pedestrian attribute datasets (the # denotes the number of).

Dataset	**# Pedestrians**	**#Attributes** (Binary/Multi-class)	**Source**
PETA	19000	61/4	outdoor & indoor
RAP	41585	69/3	indoor
RAP-2.0	84928	69/3	indoor
PA-100K	100000	26/0	outdoor
WIDER	13789	14/0	WIDER images (Xiong et al., 2015)
Market-1501	32668	26/1	outdoor
DukeMTMC	34183	23/0	outdoor
PARSE-27K	27000	8/2	outdoor
APiS	3661	11/2	KITTI (Geiger et al., 2012) , CBCL Street Scenes (Bileschi, 2006), INRIA (Dalal and Triggs, 2005) and SVS
HAT	9344	27/0	image site Flickr
CRP	27454	1/13	outdoor
CAD	1856	23/3	image site Sartorialist [2] and Flickr
BAP	8035	9/0	H3D (Bourdev and Malik, 2009) dataset PASCAL VOC 2010
UAV-Human (Li et al., 2021)	22,263	7/0	outdoor (UAV)

5.2. Evaluation Criteria

The performance of attribute classification can be evaluated with the Receiver Operating Characteristic (ROC) and the Area Under the average ROC Curve (AUC) which are calculated by two indicators, the recall rate and false positive rate. The recall rate is the fraction of the correctly detected positives over the total amount of positive samples, and the false positive rate means the fraction of the misclassified negatives out of the whole negative samples. At various threshold settings, a ROC curve can be drawn by plotting the recall rate vs. the false positive rate. However, seldom of PAR algorithms adopt these two metrics except for (Zhu et al., 2013). The Geometric Mean (G-mean) is used by (Chen et al., 2012) for the evaluation, which is a popular evaluation metric for unbalanced data classification.

In addition to aforementioned metrics, the mean accuracy (mA) is also used to evaluate the attribute recognition algorithms. For each attribute, mA calculates the classification accuracy of positive and negative samples respectively, and then gets their average values as the recognition result for the attribute. Finally, a recognition rate is obtained by taking an average over all attributes. The evaluation criterion can be calculated through the following formula:

$$mA = \frac{1}{N}\sum_{i=1}^{L}(\frac{TP_i}{P_i} + \frac{TN_i}{N_i}) \tag{1}$$

where L is the number of attributes. TP_i and TN_i are the number of correctly predicted positive and negative examples respectively, P_i and N_i are the number of positive and negative examples respectively.

Aforementioned evaluation criteria treat each attribute independently and ignore the inter-attribute correlation which exists naturally in multi-attribute recognition problem. (Li et al., 2016) named these metrics as *label-based* criteria and propose to use the *example-based* evaluation criteria inspired by a fact that example-based evaluation captures better the consistence of prediction on a given pedestrian image. Four widely used metrics, i.e., accuracy, precision, recall rate and F1 value, can be defined as:

$$Acc = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap f(x_i)|}{|Y_i \cup f(x_i)|}, \;\; Prec = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap f(x_i)|}{|f(x_i)|}, \;\; Rec = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap f(x_i)|}{|Y_i|}, \;\; F1 = \frac{2 * Prec * Rec}{Prec + Rec} \tag{2}$$

where N is the number of examples, Y_i is the ground truth positive labels of the i-th example, $f(x)$ returns the predicted positive labels for i-th example. And $|\cdot|$ means the set cardinality. Due to the ROC, AUC and G-mean are only used in a few PAR works, thus, we only report the main experimental results based on mAP, accuracy, precision, recall and F1 value in Table 2 and Table 3.

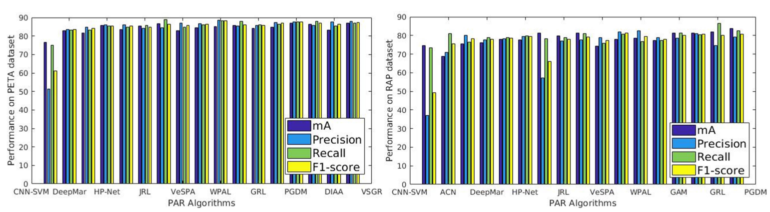

Figure 2: Comparison of selected 17 PAR algorithms (from 2014 to 2020) on the PETA and RAP dataset.

5.3. Performance Evaluation

In this section, we give a brief introduction to the performance of selected 17 PAR algorithms proposed from 2014 to 2020. As shown in Fig. 2, we can find that the baseline method CNN-SVM is outperformed by recent deep learning based PAR approaches significantly on both large scale benchmark datasets RAP and PETA. Specifically, recent deep learning approaches improve the baseline from about 50+% to 80+% on multiple evaluation metrics. These experimental results fully demonstrate the effectiveness and advantages of deep learning based PAR algorithms. Interestingly, we also find that the accuracy of current deep learning based methods are comparable, and there is no significant improvement of current methods (in 2020) compared with deep PAR algorithms proposed in several years ago. More detailed experimental results of these methods can be found in Table 2 and Table 3. Therefore, how to design new modules for the further improvement of PAR results in future works? In the following section, we propose several possible research directions for PAR.

6. Future Research Directions

More Accurate and Efficient Part Localization Algorithm Human beings could recognize the detailed attributes information in an very efficient way, because we can focus on specific regions in a glimpse and reason the attribute based on the local and global information. Therefore, it is an intuitive idea to design algorithms which can detect the local parts for accurate attribute recognition. According to section 3.2, it is easy to find that researchers are indeed more interested in mining local parts of human body. They use manual annotated or detected human body or pose information for the part localization. There are also some algorithms attempting to propose unified framework in a weakly supervised manner to jointly handle the attribute recognition and localization. We think this will also be a good and useful research direction for pedestrian attribute recognition.

Deep Generative Models for Data Augmentation In recent years, the deep generative models have made great progress and many algorithms are proposed. One intuitive research direction is how can we use deep generative models to handle the issues of low-quality person images or unbalanced data distribution? There are already many researches who focus on image generation with the guidance of text, attribute or pose information. The generated images can be used in many other tasks for data augmentation, for example, object detection, person re-identification and visual tracking (Wang et al., 2018). It is also worthy to design new algorithms to generate pedestrian images according to given attributes to augment the training data.

Table 2: An overview of PAR algorithms reviewed in this paper (Part-I).

Algorithm	Part	Attention	Seq.	C. L.	Graphic	Groups	Loss	Accuracy
Poselets (Bourdev et al., 2011) (ICCV-2011)	✓						-	*mAP* BAP/Attributes25K: 65.18/51.06
DCSA (Chen et al., 2012) (ECCV-2012)	✓				✓		SVM	-
RAD (Joo et al., 2013)(ICCV-2013)	✓						-	*mAP* HAT: 59.3
PANDA (Zhang et al., 2014) (CVPR-2014)	✓						SVM	*mAP* BAP/Attributes25K: 78.98/70.74
ACN (Sudowe et al., 2015) (ICCVW-2015)							KL-loss	*mAP* PARSE-27k: 63.6, HATDB: 66.2, BAP: 80.02
DeepSAR (Li et al., 2015) (ACPR-2015)							Softmax Loss	*Accuracy* PETA: 81.3
DeepMAR (Li et al., 2015) (ACPR-2015)							Weighted Cross-entropy Loss	*Accuracy* PETA: 82.6
MTCNN (Abdulnabi et al., 2015) (TMM-2015)						✓	Softmax Loss	*Accuracy* CAD: 92.82, AwA: 81.19
MLCNN (Zhu et al., 2015)(ICB-2015)	✓						Softmax Loss	*Accuracy* VIPeR: 74.1, GRID: 73.2
AAWP (Gkioxari et al., 2015) (ICCV-2015)	✓						SVM	*mAP* BAP: 83.6
ARAP (Luwei Yang and Tan, 2016) (BMVC-2016)	✓						Softmax Loss	*Accuracy* MPII-AlexNet: 78.00/73.2/77.74, Garment-AlexNet: 76.24/67.70/77.48
DeepCAMP (Diba et al., 2016) (CVPR-2016)	✓						Softmax Loss	*mAP* BAP: 86.6
DHC (Li et al., 2016) (ECCV-2016)	✓						Cross-entropy Loss	*mAP* BAP: 92.2, HAT: 78.0 WIDER: 81.3
PatchIt (Sudowe and Leibe, 2016) (BMVC-2016)							Cross-entropy Loss	*mAP* PARSE-27k: 72.76
HydraPlus-Net (Liu et al., 2017) (ICCV-2017)		✓					Softmax Loss	*mA/Acc/Prec/Recall/*F_1, PA-100K: 74.21/72.19/82.97/82.09/82.53, PETA: 81.77/76.13/84.92/83.24/84.07, RAP: 76.12/65.39/77.53/78.79/78.05
CAM (Guo et al., 2017) (PRL-2017)		✓					Exponential Loss	*mAP* BAP: 89.9 WIDER: 82.9
JRL (Wang et al., 2017) (ICCV-2017)	✓		✓				Cross-entropy Loss	*mA/Prec/Recall/*F_1, PETA: 85.67/86.03/85.34/85.42, RAP: 77.81/78.11/78.98/78.58
WPAL (Zhou et al., 2017) (BMVC-2017)							Weighted Cross-entropy Loss	*mA/Acc/Prec/Recall/*F_1, PETA: 85.50/76.98/84.07/85.78/84.90, RAP: 81.25/50.30/57.17/78.39/66.12
AWMT (He et al., 2017) (MM-2017)		✓		✓			Cross-entropy Loss	*mAP* CelebA: 91.80, Market-1501: 88.49, Duke: 87.53
MTCT (Dong et al., 2017) (WACV-2017)				✓		✓	t-STE Loss	*Accuracy/Precision/Recall* Street data-c: 64.35/64.97/75.66
CILICIA (Sarafianos et al., 2017) (ICCV-2017)				✓		✓	Categorical Cross-entropy Loss	*Accuracy:* SoBiR: 73.1 VIPeR: 80.5
FaFS (Lu et al., 2017) (CVPR-2017)						✓	Cross-entropy Loss	*Accuracy/Top-10 Recall* CelebA: 91.02/71.38
GAM (Fabbri et al., 2017) (AVSS-2017)	✓						Cross-entropy Loss	*mA/Acc/Prec/Rec/*F_1, RAP: 79.73/83.97/76.96/78.72/77.83
MTA-Net (Ji et al., 2020) (PRL-2020)		✓	✓				Focal Balance Loss	*mA/Acc/Prec/Rec/*F_1, RAP: 77.62/67.17/79.72/78.44/79.07, PETA: 84.62/78.80/85.67/86.42/86.04

Table 3: An overview of PAR algorithms reviewed in this paper (Part-II).

Algorithm	Part	Attention	Seq.	C. L.	Graphic	Groups	Loss	Accuracy
A-AOG (Park et al., 2018) (TPAMI-2018)	✓				✓		-	*mAP/mAC* BAP: 91.6/84.3
GRL (Zhao et al., 2018) (IJCAI-2018)	✓		✓			✓	Cross-entropy Loss	*mA/Prec/Rec/*F_1, PETA: 86.70/84.34/88.82/86.51 RAP: 81.20/77.70/80.90/79.29
LGNet (Liu et al., 2018) (BMVC-2018)	✓						Softmax Loss	*mA/Acc/Prec/Rec/*F_1, RAP: 78.68/68.00/80.36/79.82/80.09, PA-100K: 76.96/75.55/86.99/83.17/85.04
PGDM (Li et al., 2018) (ICME-2018)	✓						Weighted Cross-entropy Loss	*mA/Acc/Prec/Rec/*F_1, PETA: 82.97/78.08/86.86/84.68/85.76, RAP: 74.31/64.57/78.86/75.90/77.35, PA-100K: 74.95/73.08/84.36/82.24/83.29
DIAA (Sarafianos et al., 2018) (ECCV-2018)		✓					Weighted Focal Loss	*mA/Acc/Prec/Rec/*F_1, PETA: 84.59/78.56/86.79/86.12/86.46
VSGR (HUANG, 2019) (AAAI-2019)	✓		✓		✓		Cross-entropy Loss	*mA/Acc/Prec/Rec/*F_1, RAP: 77.91/70.04/82.05/80.64/81.34, PA-100K: 79.52/80.58/89.04/87.15/88.26, PETA: 85.21/81.25/88.43/88.42/88.42
RCRA (Xin Zhao and Yan, 2019) (AAAI-2019)		✓	✓			✓	Weighted Cross-entropy Loss	*mA/Prec/Rec/*F_1, RAP: 78.47/82.67/76.65/79.54, PETA: 85.78/85.42/88.02/86.07
IA^2-Net (Ji et al., 2019) (PRL-2019)		✓	✓			✓	Focal Cross-entropy Loss	*mA/Acc/Prec/Rec/*F_1, RAP: 77.44/67.75/79.01/77.45/78.03, PETA: 84.13/78.62/85.73/86.07/85.88
JLPLS-PAA (Tan et al., 2019) (TIP-2019)		✓					Cross-entropy Loss	*mA/Acc/Prec/Rec/*F_1, RAP: 81.25/67.91/78.56/81.45/79.98, PETA: 84.88/79.46/87.42/86.33/86.87, PA-100k: 81.61/78.89/86.83/87.73/87.27
CoCNN (Kai Han, 2019) (IJCAI-2019)	✓					✓	Cross-entropy Loss	*mA/Acc/Prec/Rec/*F_1, RAP: 81.42/68.37/81.04/80.27/80.65, PETA: 86.97/79.95/87.58/87.73/87.65, PA-100k: 80.56/78.30/89.49/84.36/86.85
DCL (Wang et al., 2019) (ICCV-2019)				✓			Cross-entropy + Triplet Loss	*mA*: RAP/CelebA:83.7/89.05
ALM (Tang et al., 2019) (ICCV-2019)	✓	✓					Weighted Binary Cross-entropy Loss	*mA/Acc/Prec/Rec/*F_1, RAP: 81.87/68.17/74.71/86.48/80.16, PETA: 86.30/79.52/85.65/88.09/86.85, PA-100k: 80.68/77.08/84.21/88.84/86.46
HAR (Wu et al., 2019) (AAAI-2020)	✓	✓				✓	Cross-entropy Loss	*mA/Acc/Prec/Rec/*F_1, RAP: 79.44/68.86/80.14/81.30/80.72, WIDER: mAP: 87.3
HFE (Yang et al., 2020) (CVPR-2020)						✓	Cross-entropy Loss & HFE Loss	Duke: 91.77, Market1501: 92.90
CAS (Zeng et al., 2020) (ICME-2020)		✓					Cross-entropy Loss	*mA/Acc/Prec/Rec/*F_1, PETA: 83.17/78.78/87.49/85.35/86.41, PA-100k: 77.20/78.09/88.46/84.86/86.62
CRM (Tan et al., 2020) (AAAI-2020)		✓					Cross-entropy Loss	*mA/Acc/Prec/Rec/*F_1, PETA: 86.96/80.38/87.81/ 87.09/87.45, RAP: 83.69/69.15/79.31/82.40/80.82, PA-100k: 82.31/79.47/87.45/87.77/87.61

Further Exploring the Visual Attention Mechanism Visual attention has drawn more and more researcher's attention in recent years. It is still one of the most popular techniques used in nowadays and integrated with every kind of deep neural networks in many tasks. Just as noted in (Mnih et al., 2014), one important property of human perception is that one does not tend to process a whole scene in its entirety at once. Instead, humans focus attention selectively on parts of the visual space to acquire information when and where it is needed, and combine information from different fixations over time to build up an internal representation of the scene, guiding future eye movements and decision making. It also substantially reduces the task complexity as the object of interest can be placed in the center of the fixation and irrelevant features of the visual environment ("clutter") outside the fixated region are naturally ignored. Designing novel attention mechanism or borrowing from other research domains for pedestrian attribute recognition maybe be an important research direction in the future.

Newly Designed Loss Functions In recent years, there are many loss functions proposed for deep neural network optimization, such as (Weighted) Cross Entropy Loss, Contrastive Loss, Center Loss, Triplet Loss, Focal Loss. Researchers also design new loss functions for the PAR, such as WPAL and AWMT, to further improving their recognition performance. It is a very important direction to study the influence of different loss functions for PAR.

Exploring More Advanced Network Architecture Existing PAR models adopts off the shelf pre-trained network on large scale dataset, as their backbone network architecture. Seldom of them consider the unique characteristics of PAR and design novel networks. Some novel networks are proposed in recent years, such as capsule network, however, there are still no attempts to use such networks for PAR. There are also works demonstrating that the deeper network architecture the better recognition performance we can obtain. Nowadays, Automatic Machine Learning solutions (AutoML) draw more and more attentions and many development tools are also released for the development, such as: AutoWEKA and Auto-sklearn. Therefore, it will be a good choice to design specific networks for person attribute recognition in future works with aforementioned approaches.

Prior Knowledge guided Learning Different from regular classification task, pedestrian attribute recognition always have its own characteristics due to the preference of human beings or natural constraints. It is an important research direction to mining the prior or common knowledge for the PAR. For example, we wear different clothes in various seasons, temperatures or occasions. On the other hand, some researchers attempt to use the history knowledge (such as: Wikipedia [3]) to help improve their overall performance. Therefore, how to use this information to explore the relations between person attributes or help the machine learning model to further understanding the attributes is still an unstudied problem.

Multi-modal Pedestrian Attribute Recognition Although existing single-modal algorithms already achieve good performance on some benchmark dataset as mentioned above. However, as is known to all, the RGB image is sensitive to illumination, bad weather (such as: rain, snow, fog), night time, *etc*. It seems impossible for us to achieve accurate pedestrian attribute recognition in all day and all weather. But the actual requirement of intelligent surveillance needs far more than this target. How can we bridge this gap? One intuitive idea is to mine useful information from other modalities, such as thermal or depth sensors, to integrate with RGB sensor. There are already many works attempt to fuse these multi-modal data and improve their final performance significantly. We think the idea of multi-modal fusion could also help improve the robustness of pedestrian attribute recognition. The thermal images can highlight the contour of human and some other wearing or carrying objects.

Video based Pedestrian Attribute Recognition Existing pedestrian attribute recognition is based on single image, however, we often obtain the video sequence captured by cameras in practical scenario. Although running existing algorithm on each video frame can be an intuitive and easy strategy, but the efficiency maybe the bottleneck for practical applications. Generally speaking, image based attribute recognition can only make use of the spatial information from the given image, which increases the difficulty of PAR due to the limited information. In contrast, given the video based PAR, we can jointly utilize the spatial and temporal information. The benefits can be listed as follows: 1). we can extend the attribute recognition into a more general case by defining more dynamic person attributes, such as “running man”; 2). the motion information can be used to reason the attributes which maybe hard to recognize in single image; 3). the general person attributes learned in videos can provide more helpful information for other video based tasks, such as video caption, video object detection. Therefore, how to recognize human attributes in practical video sequence efficiently and accurately is a problem worth studying.

Joint Learning of Attribute and Other Tasks Integrating the person attribute learning into the pipeline of other person related tasks is also an interesting and

[3]en.wikipedia.org

important research direction. There are already many algorithms proposed by considering the person attributes into corresponding tasks, such as: attribute based pedestrian detection, visual tracking, person re-identification and social activity analysis. In the future, how to better explore the fine-grained person attributes for other tasks and also use other tasks for better human attribute recognition is an important research directions.

7. Conclusion

In this paper, we give a review of PAR from traditional approaches to deep learning based algorithms in recent years. Specifically, we first introduce the background (problem formulation and challenging factors) of PAR. Then, we give a review of PAR algorithms from different perspectives, including: global based, part based, visual attention based, sequential prediction based, newly designed loss function based, curriculum learning based, graphic model based and other algorithms. After that, we discuss the specific attribute recognition, then, give a comparison between deep learning and traditional algorithm based PAR methods. After that, we show the connections between PAR and other computer vision tasks. We summarize existing benchmarks proposed for PAR, including popular datasets and evaluation criteria, and also give a brief comparison of selected 17 PAR algorithms on RAP and PETA dataset. Finally, we summarize this paper and give several possible research directions for PAR. However, due to the limited space in this paper, there are still many other works that may be related to PAR but not covered in this survey. For example, the history of the backbone deep networks used in deep PAR algorithms, the various machine learning techniques such as transfer learning, self-supervised learning, meta-learning, or active learning which may inspire the researchers to design more advanced PAR algorithms. In our future works, we will summarize these techniques which may be useful for pedestrian attribute recognition.

Acknowledgements: This work is jointly supported by Postdoctoral Innovative Talent Support Program BX20200174, China Postdoctoral Science Foundation Funded Project 2020M682828. National Nature Science Foundation of China (61976002, 62076003, 61860206004), Australian Research Council Projects FL-170100117. We also thanks all the reviewers, AE and EiC for their valuable comments and suggestions.

References

Z. Chen, W. Ouyang, T. Liu, D. Tao, A shape transformation-based dataset augmentation framework for pedestrian detection, IJCV 129 (2021) 1121–1138.

Y. Deng, P. Luo, C. C. Loy, X. Tang, Pedestrian attribute recognition at far distance, in: Proceedings of the 22nd ACM MM, 2014, pp. 789–792.

P. Sudowe, H. Spitzer, B. Leibe, Person attribute recognition with a jointly-trained holistic cnn model, in: IEEE ICCV Workshops, 2015, pp. 87–95.

D. Li, X. Chen, K. Huang, Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios, in: ACPR, IEEE, 2015, pp. 111–115.

A. H. Abdulnabi, G. Wang, J. Lu, K. Jia, Multi-task cnn model for attribute prediction, IEEE TMM 17 (2015) 1949–1959.

H. S. C. L. C. X. C. X. Kai Han, Yunhe Wang, Attribute aware pooling for pedestrian attribute recognition, in: IJCAI, 2019.

L. Bourdev, S. Maji, J. Malik, Describing people: A poselet-based approach to attribute classification, in: IEEE ICCV, 2011, pp. 1543–1550.

J. Joo, S. Wang, S.-C. Zhu, Human attribute recognition by rich appearance dictionary, in: IEEE ICCV, 2013, pp. 721–728.

S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: IEEE CVPR, volume 2, 2006, pp. 2169–2178.

N. Zhang, M. Paluri, M. Ranzato, T. Darrell, L. Bourdev, Panda: Pose aligned networks for deep attribute modeling, in: IEEE CVPR, 2014, pp. 1637–1644.

G. Gkioxari, R. Girshick, J. Malik, Actions and attributes from wholes and parts, in: IEEE ICCV, 2015.

R. Girshick, F. Iandola, T. Darrell, J. Malik, Deformable part models are convolutional neural networks, in: IEEE CVPR, 2015, pp. 437–446.

J. Zhu, S. Liao, D. Yi, Z. Lei, S. Z. Li, Multi-label cnn based pedestrian attribute learning for soft biometrics, in: IEEE ICB, 2015, pp. 535–540.

C. Tang, L. Sheng, Z. Zhang, X. Hu, Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization, in: IEEE ICCV, 2019, pp. 4997–5006.

Y. W. S. L. Luwei Yang, Ligeng Zhu, P. Tan, Attribute recognition from adaptive parts, in: BMVC, 2016, pp. 81.1–81.11. doi:10.5244/C.30.81.

A. Diba, A. Mohammad Pazandeh, H. Pirsiavash, L. Van Gool, Deepcamp: Deep convolutional action & attribute mid-level patterns, in: IEEE CVPR, 2016, pp. 3557–3565.

D. Li, X. Chen, Z. Zhang, K. Huang, Pose guided deep model for pedestrian attribute recognition in surveillance scenarios, in: IEEE ICME, 2018, pp. 1–6.

Y. Li, C. Huang, C. C. Loy, X. Tang, Human attribute recognition by deep hierarchical contexts, in: ECCV, Springer, 2016, pp. 684–700.

P. Liu, X. Liu, J. Yan, J. Shao, Localization guided learning for pedestrian attribute recognition, in: BMVC, 2018.

X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, X. Wang, Hydraplus-net: Attentive deep features for pedestrian analysis, in: IEEE ICCV, 2017, pp. 350–359.

M. S. Sarfraz, A. Schumann, Y. Wang, R. Stiefelhagen, Deep view-sensitive pedestrian attribute inference in an end-to-end model, arXiv:1707.06089 (2017).

N. Sarafianos, X. Xu, I. A. Kakadiaris, Deep imbalanced attribute classification using visual attention aggregation, in: ECCV, Springer, 2018, pp. 708–725.

H. Guo, X. Fan, S. Wang, Human attribute recognition by refining attention heat map, Pattern Recognition Letters 94 (2017) 38–45.

Z. Tan, Y. Yang, J. Wan, H. Wan, G. Guo, S. Z. Li, Attention based pedestrian attribute analysis, IEEE TIP (2019).

Z. Ji, E. He, H. Wang, A. Yang, Image-attribute reciprocally guided attention network for pedestrian attribute recognition, Pattern Recognition Letters 120 (2019) 89–95.

M. Wu, D. Huang, Y. Guo, Y. Wang, Distraction-aware feature learning for human attribute recognition via coarse-to-fine attention mechanism, arXiv:1911.11351 (2019).

H. Zeng, H. Ai, Z. Zhuang, L. Chen, Multi-task learning via co-attentive sharing for pedestrian attribute recognition, in: ICME, IEEE, 2020, pp. 1–6.

S. Zhang, Z. Song, X. Cao, H. Zhang, J. Zhou, Task-aware attention model for clothing attribute prediction, TCSVT 30 (2019) 1051–1064.

J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, W. Xu, Cnn-rnn: A unified framework for multi-label image classification, in: IEEE CVPR, 2016, pp. 2285–2294.

J. Wang, X. Zhu, S. Gong, W. Li, Attribute recognition by joint recurrent learning of context and correlation, in: IEEE ICCV, 2017, pp. 531–540.

X. Zhao, L. Sang, G. Ding, Y. Guo, X. Jin, Grouping attribute recognition for pedestrian with joint recurrent learning., in: IJCAI, 2018, pp. 3177–3183.

H. Liu, J. Wu, J. Jiang, M. Qi, R. Bo, Sequence-based person attribute recognition with joint ctc-attention model, arXiv preprint arXiv:1811.08115 (2018).

G. D. J. H. N. D. Xin Zhao, Liufang Sang, C. Yan, Recurrent attention model for pedestrian attribute recognition, in: AAAI, 2019.

Y. Zhou, K. Yu, B. Leng, Z. Zhang, D. Li, K. Huang, B. Feng, C. Yao, et al., Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization, in: BMVC, 2017.

K. He, Z. Wang, Y. Fu, R. Feng, Y.-G. Jiang, X. Xue, Adaptively weighted multi-task deep network for person attribute classification, in: ACM MM, 2017, pp. 1636–1644.

E. Yaghoubi, D. Borza, J. Neves, A. Kumar, H. Proença, An attention-based deep learning model for multiple pedestrian attributes recognition, arXiv preprint arXiv:2004.01110 (2020).

J. Yang, J. Fan, Y. Wang, Y. Wang, W. Gan, L. Liu, W. Wu, Hierarchical feature embedding for attribute recognition, in: IEEE CVPR, 2020, pp. 13055–13064.

J. Jia, H. Huang, W. Yang, X. Chen, K. Huang, Rethinking of pedestrian attribute recognition: Realistic datasets with efficient method, arXiv preprint arXiv:2005.11909 (2020).

Z. Ji, Z. Hu, E. He, J. Han, Y. Pang, Pedestrian attribute recognition based on multiple time steps attention, Pattern Recognition Letters (2020).

Q. Dong, S. Gong, X. Zhu, Multi-task curriculum transfer deep learning of clothing attributes, in: IEEE WACV, 2017, pp. 520–529.

N. Sarafianos, T. Giannakopoulos, C. Nikou, I. A. Kakadiaris, Curriculum learning for multi-task classification of visual attributes, in: IEEE ICCV, 2017, pp. 2608–2615.

J. Zhu, S. Liao, Z. Lei, S. Z. Li, Multi-label convolutional neural network based pedestrian attribute classification, IVC 58 (2017) 224–229.

N. Sarafianos, T. Giannakopoulos, C. Nikou, I. A. Kakadiaris, Curriculum learning of visual attribute clusters for multi-task classification, Pattern Recognition 80 (2018) 94–108.

Y. Wang, W. Gan, W. Wu, J. Yan, Dynamic curriculum learning for imbalanced data classification, ICCV (2019).

H. Chen, A. Gallagher, B. Girod, Describing clothing by semantic attributes, in: ECCV, Springer, 2012, pp. 609–623.

S. Park, B. X. Nie, S.-C. Zhu, Attribute and-or grammar for joint parsing of human pose, parts and attributes, IEEE TPAMI 40 (2018) 1555–1569.

Q. L. X. Z. R. H. K. HUANG, Visual-semantic graph reasoning for pedestrian attribute recognition, in: AAAI, 2019.

Z. Tan, Y. Yang, J. Wan, G. Guo, S. Z. Li, Relation-aware pedestrian attribute recognition with graph convolutional networks., in: AAAI, 2020, pp. 12055–12062.

W. Wang, Y. Xu, J. Shen, S.-C. Zhu, Attentive fashion grammar network for fashion landmark detection and clothing category classification, in: IEEE CVPR, 2018, pp. 4271–4280.

P. Sudowe, B. Leibe, Patchit: Self-supervised network weight initialization for fine-grained recognition., in: BMVC, 2016.

Y. Lu, A. Kumar, S. Zhai, Y. Cheng, T. Javidi, R. Feris, Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification, in: CVPR, volume 1, 2017, p. 6.

M. Fabbri, S. Calderara, R. Cucchiara, Generative adversarial models for people attribute recognition in surveillance, in: IEEE AVSS, 2017, pp. 1–6.

G. D. J. H. L. L. Liuyu Xiang, Xiaoming Jin, Incremental few-shot learning for pedestrian attribute recognition, in: IJCAI, 2019.

M. Mirza, S. Osindero, Conditional generative adversarial networks, Manuscript: https://arxiv.org/abs/1709.02023 (2014).

Y. Zhang, P. Zhang, C. Yuan, Z. Wang, Texture and shape biased two-stream networks for clothing classification and attribute recognition, in: IEEE CVPR, 2020, pp. 13538–13547.

J. Jia, H. Huang, X. Chen, K. Huang, Rethinking of pedestrian attribute recognition: A reliable evaluation under zero-shot pedestrian identity setting, arXiv preprint arXiv:2107.03576 (2021).

X. Zheng, Y. Guo, H. Huang, Y. Li, R. He, A survey to deep facial attribute analysis, arXiv preprint arXiv:1812.10265 (2018).

B. Fasel, J. Luettin, Automatic facial expression analysis: a survey, Pattern recognition 36 (2003) 259–275.

P. Rodríguez, G. Cucurull, J. M. Gonfaus, F. X. Roca, J. Gonzalez, Age and gender recognition in the wild with deep attention, Pattern Recognition 72 (2017) 563–571.

K. Li, J. Xing, W. Hu, S. J. Maybank, D2c: Deep cumulatively and comparatively learning for human age estimation, Pattern Recognition 66 (2017) 95–105.

J. Xing, K. Li, W. Hu, C. Yuan, H. Ling, Diagnosing deep learning models for high accuracy age estimation from a single image, Pattern Recognition 66 (2017) 106–116.

G. Antipov, M. Baccouche, S.-A. Berrani, J.-L. Dugelay, Effective training of convolutional neural networks for face-based gender and age prediction, Pattern Recognition 72 (2017) 15–26.

H. Liu, J. Lu, J. Feng, J. Zhou, Group-aware deep feature learning for facial age estimation, Pattern Recognition 66 (2017) 82–94.

K. Chen, K. Jia, Z. Zhang, J.-K. Kämäräinen, Spectral attribute learning for visual regression, Pattern Recognition 66 (2017) 74–81.

A. Hadid, M. Pietikäinen, Combining appearance and motion for face and gender recognition from videos, Pattern Recognition 42 (2009) 2818–2827.

A. Branca, M. Leo, G. Attolico, A. Distante, Detection of objects carried by people, in: IEEE ICIP, volume 3, 2002.

F. Ghadiri, R. Bergevin, G.-A. Bilodeau, From superpixel to human shape modelling for carried object detection, Pattern Recognition 89 (2019) 134–150.

D. Damen, D. Hogg, Detecting carried objects from sequences of walking pedestrians, IEEE TPAMI 34 (2011) 1056–1067.

F. Ghadiri, R. Bergevin, G.-A. Bilodeau, Carried object detection based on an ensemble of contour exemplars, in: ECCV, Springer, 2016, pp. 852–866.

Y. Tian, P. Luo, X. Wang, X. Tang, Pedestrian detection aided by deep learning semantic tasks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5079–5087.

Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, Y. Yang, Improving person re-identification by attribute and identity learning, Pattern Recognition (2019).

K. Han, J. Guo, C. Zhang, M. Zhu, Attribute-aware attention model for fine-grained representation learning, in: ACM MM, 2018, pp. 2040–2048.

C. Su, S. Zhang, J. Xing, W. Gao, Q. Tian, Deep attributes driven multi-camera person re-identification, in: ECCV, Springer, 2016, pp. 475–491.

S. Khamis, C.-H. Kuo, V. K. Singh, V. D. Shet, L. S. Davis, Joint learning for attribute-consistent person re-identification, in: ECCV, Springer, 2014, pp. 134–146.

A. Li, L. Liu, K. Wang, S. Liu, S. Yan, Clothing attributes assisted person reidentification, IEEE TCSVT 25 (2015) 869–878.

R. Layne, T. M. Hospedales, S. Gong, Towards person identification and re-identification with attributes, in: ECCV, Springer, 2012, pp. 402–412.

R. Layne, T. M. Hospedales, S. Gong, Attributes-based re-identification, in: Person Re-Identification, Springer, 2014, pp. 93–117.

R. Layne, T. M. Hospedales, S. Gong, Q. Mary, Person re-identification by attributes., in: BMVC, volume 2, 2012, p. 8.

A. Schumann, R. Stiefelhagen, Person re-identification by deep learning attribute-complementary information, in: IEEE CVPR Workshops, 2017, pp. 20–28.

S. Li, H. Yu, W. Huang, J. Zhang, Attributes-aided part detection and refinement for person re-identification, arXiv preprint arXiv:1902.10528 (2019).

H. Ling, Z. Wang, P. Li, Y. Shi, J. Chen, F. Zou, Improving person re-identification by multi-task learning, Neurocomputing 347 (2019) 109–118.

C. Su, S. Zhang, F. Yang, G. Zhang, Q. Tian, W. Gao, L. S. Davis, Attributes driven tracklet-to-tracklet person re-identification using latent prototypes space mapping, Pattern Recognition 66 (2017) 4–15.

Y. Chen, S. Duffner, A. Stoian, J.-Y. Dufour, A. Baskurt, Deep and low-level feature based attribute learning for person re-identification, IVC 79 (2018) 25–34.

X. Wang, T. Zhang, D. R. Tretter, Q. Lin, Personal clothing retrieval on photo collections by color and attributes, IEEE TMM 15 (2013) 2035–2045.

Q. Chen, J. Huang, R. Feris, L. M. Brown, J. Dong, S. Yan, Deep domain adaptation for describing people based on fine-grained clothing attributes, in: IEEE CVPR, 2015, pp. 5315–5324.

M. Ziaeefard, R. Bergevin, Semantic human activity recognition: A literature review, Pattern Recognition 48 (2015) 2329–2345.

J. Liu, B. Kuipers, S. Savarese, Recognizing human actions by attributes, in: IEEE CVPR, 2011, pp. 3337–3344.

E. Murphy-Chutorian, M. M. Trivedi, Head pose estimation in computer vision: A survey, IEEE TPAMI 31 (2008) 607–626.

L. Huang, J. Peng, R. Zhang, G. Li, L. Lin, Learning deep representations for semantic image parsing: a comprehensive overview, Frontiers of Computer Science 12 (2018) 840–857.

W. Xiao, Z. Shaofei, Y. Rui, L. Bin, T. Jin, Pedestrian attribute recognition: A survey, arXiv:1901.07474 (2019).

Y. Xiong, K. Zhu, D. Lin, X. Tang, Recognize complex events from static images by fusing deep channels, in: CVPR, 2015, pp. 1600–1609.

A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? the kitti vision benchmark suite, in: CVPR, 2012, pp. 3354–3361.

S. M. Bileschi, StreetScenes: Towards scene understanding in still images, Technical Report, MASSACHUSETTS INST OF TECH CAMBRIDGE, 2006.

N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: CVPR, volume 1, 2005, pp. 886–893.

L. Bourdev, J. Malik, Poselets: Body part detectors trained using 3d human pose annotations, in: ICCV, 2009, pp. 1365–1372.

T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, Z. Li, Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles, in: IEEE CVPR, 2021, pp. 16266–16275.

J. Zhu, S. Liao, Z. Lei, D. Yi, S. Li, Pedestrian attribute classification in surveillance: Database and evaluation, in: IEEE ICCV Workshops, 2013, pp. 331–338.

D. Li, Z. Zhang, X. Chen, H. Ling, K. Huang, A richly annotated dataset for pedestrian attribute recognition, arXiv:1603.07054 (2016).

X. Wang, C. Li, B. Luo, J. Tang, Sint++: Robust visual tracking via adversarial positive instance generation, in: IEEE CVPR, 2018, pp. 4864–4873.

V. Mnih, N. Heess, A. Graves, et al., Recurrent models of visual attention, in: NIPS, 2014, pp. 2204–2212.