A Realistic Predictor for Pedestrian Attribute Recognition

Andreas Specker

Vision and Fusion Laboratory
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT), Germany
andreas.specker@kit.edu

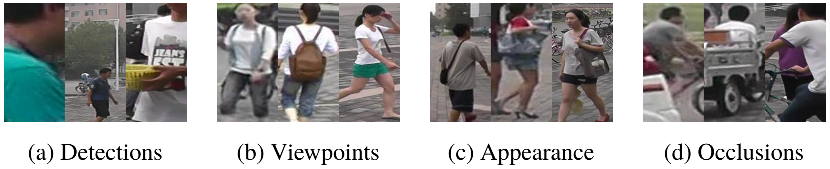
Technical Report IES-2019-11

Abstract

The application of video surveillance systems in public areas to ensure public security is becoming increasingly important. A major task when evaluating the arising amount of video data is to find the occurrences of a person-of-interest on the basis of a testimony. For the comparison of a person's description with persons in the video data, the attributes of all persons must be recognized automatically. However, typical approaches to pedestrian attribute recognition simply predict all attributes for a person, regardless the visibility of relevant attributes. To address this problem, the concept of realistic predictors is used in this work to determine and improve the reliability of pedestrian attribute recognition.

1 Introduction

Nowadays, more and more video surveillance systems are used to ensure public security. Due to the large amount of image and video footage that is recorded by



(a) Detections (b) Viewpoints (c) Appearance (d) Occlusions

Figure 1.1: Different challenges in recognizing pedestrian attributes. Poor detections and occlusions can lead to only partially visible persons in images. Moreover, some attributes like backpack may not be visible from all point of views and attributes such as handbags may appear as many different types.

such systems, manual evaluation is hardly possible, which is why intelligent and automatic analysis systems are required. One of the most important evaluation tasks that can be automatically solved by applying convolutional neural networks (CNN) is person re-identification which aims to find all occurrences of a person-of-interest in the data. Typically, such a search is performed based on a cropped image of the person the system operators are interested in. But since it is not possible to cover all areas with CCTV cameras, one cannot be sure that a query image of the person-of-interest is always available. Thus, in such cases, descriptions of the semantic attributes are the only clues on which the person search can be based. The query attributes can be easily and directly extracted from witness descriptions. In order to find all persons corresponding to the obtained attributes, the semantic attributes of the persons present in the surveillance material must be recognized.

This pedestrian attribute recognition in an uncooperative, real-world scenario suffers from a lot of different challenges. Some of the most severe issues to overcome are visualized in Figure 1.1. Stable recognition of a person's semantic attributes is only possible if clean cutouts are available. But sometimes person detectors provide bad detections which show a lot of background clutter or only parts of a human body. Moreover, the view angle is an factor that greatly influences the appearance of a person. Attributes as for instance backpack may not be visible from every point of view. Similar issues arise from occlusions which make it difficult or impossible to determine certain attributes. Lastly, attributes, such as handbag in Figure 1.1(c), can differ greatly regarding their appearance. Handbags come in different sizes and colors making the recognition task harder.

All those challenges indicate that meaningful attribute predictions can not be given in all cases. If, for instance, the lower-body of a person is occluded by a vehicle, no well-founded statement about the length of the lower-body clothing can be made. Although this is a very important topic, it is not present in existing literature regarding to pedestrian attribute recognition. However, with regard to typical one-hot classification, Wang et al. [17] present an approach which takes into account the hardness of the input images and only provides classification results if a reliable estimation is possible. Since attribute recognition, albeit multi-class, is a classification problem as well, the core idea of this work is to transfer and adapt the concept of realistic predictors to this task.

2 Related work

Generally, pedestrian attribute recognition approaches from related literature can be roughly divided into three different categories: global, part-based and attention-based methods.

Global Models Especially early deep learning-based works on pedestrian attribute recognition predict semantic attributes on solely a whole body image of a person. In [16] for instance, a multi-branch architecture is applied that contains a separate classification layer and loss for each attribute. In contrast, some works showed that it is advantageous not to learn all the attributes separately but instead learn them all together [7] or partitioned in groups of corresponding attributes [1]. In addition to that, the authors in [7] propose to weight the attributes during loss calculation according to their frequency of occurrence in the dataset to handle the large imbalances of attribute values. The results of newer works [15], however, indicate that with the development of larger CNN models the joint learning of attributes is not always beneficial and higher accuracies can be achieved if separate networks are used for different attributes. In general, global models are simple and therefore very efficient compared to more complex architectures. These results in faster training and testing, though only using coarse information. Differences between global attributes, as gender, and small-scale attributes such as shoes or glasses are not taken into account and aggravate the recognition task.

Attention-based Models Attention-based methods aim to guide the network to focus on the most important regions of activation maps or features. [12, 13] propose networks that are capable of implicitly learning visual attention maps. A special feature of [12] is the use of a multi-directional attention mechanism which means that attention is shared between different semantic layers of the network. Moreover, Sarfraz et al. [14] introduce an approach to learn view-sensitive embeddings since the viewpoint of a person is really important with respect to the appearance of attributes. To improve attention maps explicitly, in [5] attention maps are refined using a exponential loss function. Although some attention-based methods are proposed in literature, the gain in accuracy is still limited compared to other research fields such as for instance person re-identification.

Part-based Models Part-based algorithms jointly leverage local and global information to improve recognition accuracy. This is done by either localizing body parts of persons using an extern [4, 9] or intern [3, 11, 18] module. In [4] patches obtained from a part detector are fed into a fine-grained classification model. Similar to that, [10] proposes to use the detector features of the whole person and detected parts as input patches for attribute classification layers. A slightly different way is followed in [9]. Instead of bounding boxes estimated by a body part detector, pose key points are exploited to localize meaningful body part regions. In contrast to these approaches, [18] introduces a method by which part localization and attribute classification is jointly learned in an end-to-end manner. In [3], the authors use mid-level image patches as representations of human body parts. Moreover, LGNet is presented in [11]. Consisting of a global and a local network branch, part detection is performed by creating so-called EdgeBoxes that are applied in a Region-of-Interest pooling module. Such part-based models are less efficient compared to simple global models but instead are able to focus on fine-grained information which is very important for recognition of very local attribute, as for instance glasses or shoes. However, it is important that body parts can be accurately detected because otherwise the approaches suffer from focusing on irrelevant regions of the input image.

Although part-based models implicitly handle the visibility of body parts or attributes, none of the approaches in literature deal with the fact that in a uncooperative real-world scenario attributes cannot be predicted for imperfect person image crops or occluded body parts. Therefore this work aims to close this research gap by investigating the concept of realistic predictors which is detailed in the following.

3 Methods

In this chapter, the baseline classification model is presented followed by a detailed description of the realistic predictor approach.

3.1 Baseline model

The baseline model is based on the typical classification pipeline for global pedestrian attribute recognition. Images are pre-processed and data augmentation is performed. Afterwards, images are fed into a backbone network with appended fully-connected classification layer and output probabilities are computed using the sigmoid function. In this case, the task is considered a multi-class classification task which means that all attributes are simultaneously predicted using a single classification layer. Sigmoid cross-entropy loss function (SCEL) is applied to train the CNN model. To handle the imbalanced distribution of positive attribute labels in the dataset, a weighting factor is added to the loss computation as proposed in [7]. Let $y_i^c \in [0,1]$ be the target label of the cth attribute of the ith sample and p^c the positive ratio of this attribute in the dataset. Then the weighting factor w_i^c can be computed independently for each attribute and input image as follows:

$$w_i^c = \begin{cases} \exp(\frac{(1-p^c)}{\sigma^2}) & \text{, if } y_i^c = 1 \\ \exp(\frac{p^c}{\sigma^2}) & \text{, if } y_i^c = 0 \end{cases} \tag{3.1}$$

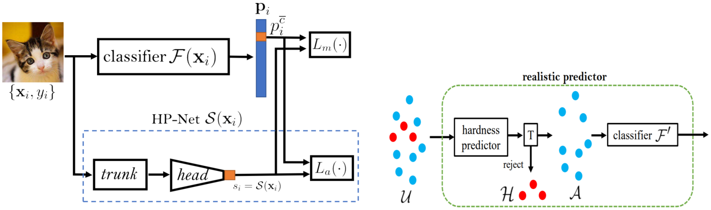


Figure 3.1: The general idea of realistic predictors. On the left, the architecture is shown consisting of two branches: a classification and a hardness prediction one respectively. The figure on the right depicts the testing stage. Samples with a hardness score above a threshold T are discarded and not fed into the classifier. [17]

σ stands for a hyperparameter which is set to 1 in all experiments. This weighting factor ensures that the network focuses on rare attributes by increasing the weight of such samples.

3.2 Realistic predictors

The concept of realistic predictors is adapted from [17]. The general approach is visualized in Figure 3.1. A network with two branches was designed to simultaneously train a classifier and a so-called hardness predictor. The classifier outputs probabilities p_i for each class whereas the hardness prediction network computes hardness scores. Hardness scores s_i are understood as predictions of the difficulty of the classification task for a specific input image. So, for instance, the hardness score should be higher if an object is only partially visible in comparison with a clean cut of the object of interest. The testing protocol is visualized in Figure 3.1 on the right. First, the hardness for all samples is predicted. To find those images for which no reliable classification can be provided, hard samples are discarded based on a threshold T. The remaining samples are then forwarded through the classifier and a class prediction is produced. In practice, only attributes for which the classifier is certain would be output and then used for person retrieval.

Two different losses are used to train the two network branches. For training the classifier, the use of a weighted softmax cross-entropy loss function is proposed. This loss function L_m is shown in the following equation where N stands for the number of samples in the batch and $p_i^{\bar{c}}$ depicts the predicted probability for target class $\bar{c}$ and sample i.

$$L_m = -\sum_{i=1}^{N} s_i \log p_i^{\bar{c}} \tag{3.2}$$

As mentioned earlier, the original paper deals with a one-hot classification problem in contrast to the pedestrian attribute task. Persons have several attributes at the same time, like a woman wearing a red shirt and blue jeans, and thus multiple classes can be true. Therefore the loss function for the multi-class task is adapted as follows:

$$L_m = -\sum_{i=1}^{N} \sum_{c=1}^{C} \left[y_i^c \log p_i^c + (1 - y_i^c) \log(1 - p_i^c) \right] \tag{3.3}$$

In addition to the sum over all samples, the sum of cross-entropy losses for all attributes is computed. C denotes the number of different semantic attributes in this case and $y_i^c \in [0, 1]$ is the target label of the cth attribute.

Another alteration that was made is that the feedback of the predicted hardness score s_i is omitted in contrast to the original paper. Whereas the authors propose this term to focus on those samples that are particularly hard during training, this is not necessarily beneficial for attribute recognition. In the object classification approach one can be certain that the object is actually present and visible in the input image. In contrast, especially small-scale attributes are often occluded and therefore not visible which could lead to a decrease in recognition accuracy if such samples are preferred during the training process. The network would not be able to base its decision on meaningful clues and to learn important information.

For training the hardness predictor, another loss function is proposed in [17].

$$L_a = -\sum_{i=1}^{N} \left[p_i^{\bar{c}} \log (1 - s_i) + \left(1 - p_i^{\bar{c}}\right) \log s_i \right] \tag{3.4}$$

The goal of this function is to produce large hardness scores if and only if the cross-entropy loss of the classification branch is high and vice versa. Therefore, a kind of inverse cross-entropy loss is used. The loss function gets minimal if $s_i = 1 - p_i^{\bar{c}}$ applies. In words, the hardness score is forced to be equal to the classification error measured by the prediction probability. Moreover, the more the estimated class probability differs from the target value the higher the loss of the hardness predictor.

Analogous to the classification loss function, the hardness predictor loss calculation has also be modified to match the requirements of the multi-class attribute classification problem. Again, the loss function is expanded to consider each attribute. Since in contrast to the one-class classification problem not only one positive class is relevant but instead the presence as well as the absence of all attributes, loss calculation is also based on the target label, as can be seen in the equation hereafter.

$$L_a = -\sum_{i=1}^{N}\sum_{c=1}^{C}\left[\Delta p_i^c \log\left(1 - s_i^c\right) + \left(1 - \Delta p_i^c\right)\log s_i^c\right], \tag{3.5}$$

$$with \Delta p_i^c = |y_i - p_i^c| \tag{3.6}$$

Thereby, the hardness predictor learns to estimate the difficulty of an image regardless of an attribute being present or not in the training image. This is ensured by applying the absolute value of the difference between the target class label y_i^c and the predicted probability of the presence of an attribute p_i^c instead of using p_i^c directly.

Since the training of the hardness predictor network also suffers from data imbalances, DeepMAR weighting can be applied here as well, thus reducing the influence of unbalanced attributes distribution on the training.

3.3 Determination of thresholds

To improve the accuracy of pedestrian attribute recognition, meaningful thresholds for hardness scores need to be determined. It is important to find a good trade-off between improving accuracy and rejecting as few samples as possible. Thus, multiple strategies to seek for meaningful thresholds are proposed and compared in the evaluation chapter. The thresholds are computed for each attribute independently making use of the evaluation data. To avoid that too much samples of an attribute are discarded, optimization is stopped as soon as the threshold is below that of the quantile rejection method.

Threshold rejection As a baseline for comparison of the other rejection approaches, one single threshold which is applied to all attributes is determined.

Quantile rejection In contrast, quantile rejection method sets the thresholds to a value so that a predefined portion of validation samples is discarded. Since the distribution of the hardness scores may vary between validation and testing data, the proportion of rejected samples can differ during testing stage.

Mean accuracy / F1 rejection This rejection approach aims to optimize the target evaluation metric, either mean accuracy or F1 score. The threshold value is lowered until the mean accuracy no longer increases or until the stop criterion mentioned above is reached.

4 Evaluation

The previously introduced approaches are evaluated and discussed in the following. After some details about the datasets used and the experimental setup, the results of the experiments are presented.

4.1 Datasets

The experiments are conducted on two different publicly available datasets. Both datasets contain person bounding boxes that are all taken from videos of surveillance cameras. A brief introduction to RAP-2.0 and PA-100K datasets is given in the following. Some sample images of both dataset can be found in Figure 4.1.

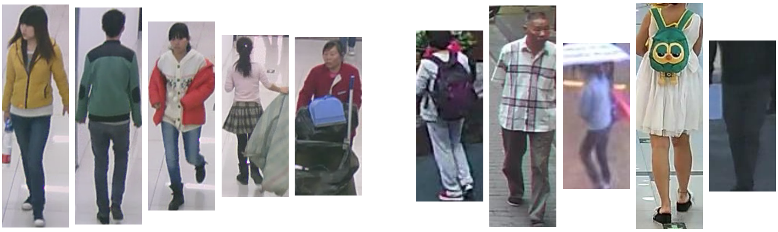


Figure 4.1: Randomly selected images from the datasets are shown for comparison. Figures (a) - (e) are taken from the RAP-2.0 dataset whereas Figures (f) - (j) are from the PA-100K dataset.

The **RAP-2.0** [8] dataset consists of 84,928 images taken from 26 different cameras. All cameras were mounted indoor and show scenes of a shopping mall. 72 different binary attributes ranging from gender to attachments are annotated. Since the distributions of the attribute annotations are highly unbalanced, only those attributes with a positive ratio greater than 1 % are used in the experiments. After discarding very rare attributes, 54 attributes remain whose positive ratios are shown in Figure 4.2.

Unlike the RAP-2.0 dataset, the **PA-100K** [12] dataset contains images recorded in an outdoor setting. According to the dataset name, 100,000 images from 598 different cameras are included and 26 binary attributes are provided. Moreover, distributions of attribute annotations are more balanced.

4.2 Experimental setup

Data pre-processing and augmentation During training phase, images are resized and randomly cropped to match the input size of the CNN. In addition, random flipping is applied to increase the diversity of training data.

Backbone model Experiments with different backbone models were carried out. Since the observations presented in this chapter are valid regardless of the CNN model used, only results for ResNet-50 [6] are presented and discussed.

Parameters To train the models, a multi-step scheduling scheme was applied in all experiments. Two steps are performed with a decay factor of 0.1. RAP-2.0

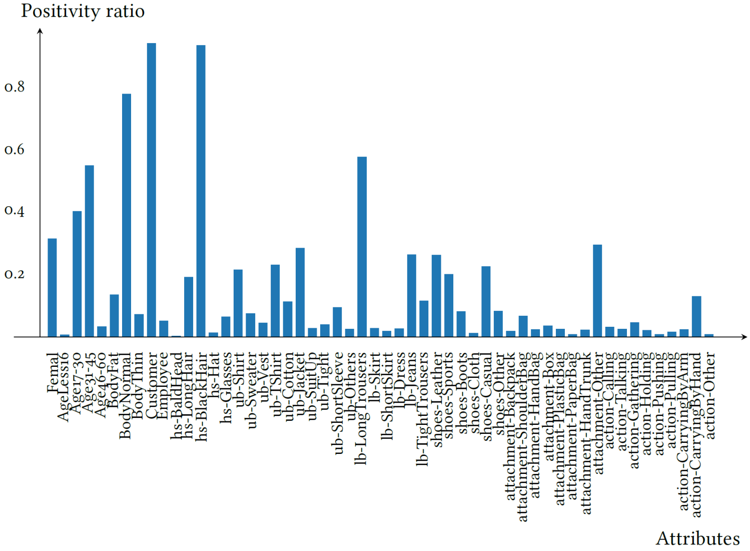


Figure 4.2: Positive ratios of RAP-2.0 attributes. Only few attributes have balanced distributions while most attributes such as *attachment-backpack* occur very rarely.

models were trained for a total of 180 epochs with steps after 60 and 120 epochs. The learning rate for the Adam optimizer was initially set to 10^{-4} for the classifier and 10^{-5} for the hardness predictor, respectively. For training the networks with the PA-100K dataset, parameters were set to the values suggested in [2].

4.3 Hardness prediction

Table 4.1 presents the attribute recognition results of the classifiers. Using positive ratio-based DeepMAR weighting of the loss during training significantly increases the recognition performance by reducing the influence of imbalanced attribute distributions. Moreover, the results clearly indicate that using feedback of the HP-Net for training the classifier network is not beneficial for pedestrian

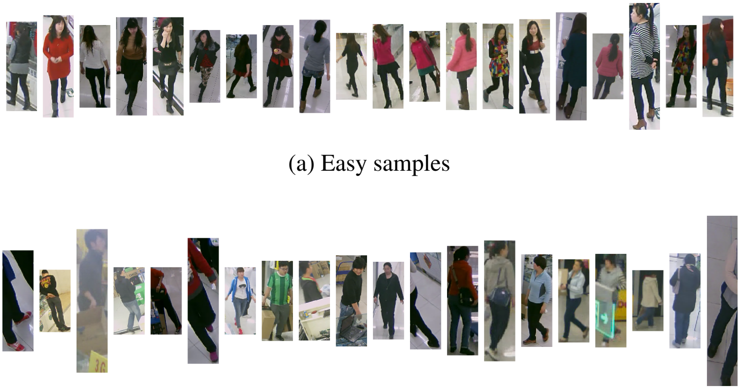
Table 4.1: Quantitative evaluation of baseline methods on RAP-2.0 dataset. DeepMAR weighting of training samples greatly improves mA. Training the classifier with HP-Net feedback deteriorates the results in all metrics.

Model	mA	Acc	Prec	Rec	F1
SCEL	64.29	62.26	**82.55**	70.09	75.81
DeepMAR	**73.05**	**63.99**	77.01	**77.17**	**77.09**
SCEL + HP-Net feedback	61.93	52.00	69.33	66.81	68.05
DeepMAR + HP-Net feedback	67.32	61.18	76.74	73.49	75.08

attribute recognition. In the original approach this feedback was proposed to force the classifier to focus on those samples which are hard to classify. But in contrast to typical image classification, attributes are small-scale features and thus not necessarily visible in hard-to-classify images. As a results, focusing on such hard samples confuses the CNN and accuracy decreases regarding all metrics as can be seen from the experimental results in the table.

Next, it is important to evaluate the quality of the given hardness predictor. For this purpose, Figures 4.3 and 4.4 show person images assessed as easy as well as hard are displayed. Figure 4.3 visualizes samples for the gender attribute. The qualitative results seem reasonable. It is easy for the classifier to classify a person as a woman if the person is wearing a skirt or has long hair that is clearly visible. In contrast, hard samples are images showing only partial persons such as the first image in Figure 4.3(b). Also a human cannot make a reliable statement about the sex, because only the legs of the person are visible. Moreover, images on which the length of the hair is not clearly visible are hard to assess for the classifier and therefore more prone to misclassification.

These observations are valid for many of the attributes but there are attributes, like backpack, for which different results are received. As an example, easy and hard samples for the attribute *Backpack* are shown in Figure 4.4. All easy samples show persons without a backpack whereas each of the persons from the particularly hard samples wears a backpack. So, in this case it seems that the decision between hard and easy images is only taken based on the presence of



(a) Easy samples

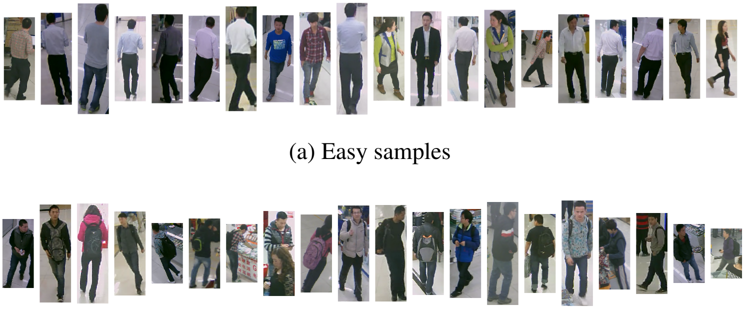
(b) Hard samples

Figure 4.3: Hard and easy samples for the attribute *Gender* of the RAP-2.0 dataset based on the estimated hardness scores. Samples that are considered easy or hard appear to be reasonable for this attribute.

the attribute and by that equals the classifier instead of providing independent hardness predictions. This indicates that, albeit the hardness predictor loss is weighted by the positive ratio of attributes, the imbalance of attributes in the training data still plays a big role and influences the recognition accuracy negatively. Since only about 1 % of the training images show persons with backpacks, the network can achieve good results by only predicting no backpack. Thus, the loss gets minimal for such images and high for images with backpacks. As a result, the hardness predictor learns to discriminate between the values of the attribute and not to predict the hardness of the attribute recognition task.

4.4 Realistic prediction

Based on the finding that the hardness predictor can give meaningful estimates of the degree of difficulty of samples, the realistic predictor can be formed by combining the classifier with a hardness-based rejection. Table 4.2 presents



(a) Easy samples

(b) Hard samples

Figure 4.4: Hard and easy samples for the attribute *Backpack* of the RAP-2.0 dataset based on the estimated hardness scores. In contrast to *Gender*, persons with backpacks are considered hard-to-classify due to the high attribute imbalance.

the results for different rejection strategies and compares them to confidence score-based rejection. Improvements in instance-based metrics can be observed, independent of the applied rejection method. The mA-score decreases except for the mA rejection. This is due to the side effects of unbalanced attributes, which are always predicted as false and thus reach only a minimum mA score of 0.5. When comparing rejection methods, threshold strategy achieves the best F1 scores whereas, as mentioned above, mA rejection leads to highest mA results. Although hardness prediction-based rejection of attributes increases the performance, rejection on the basis of class probabilities achieves similar or even better performance, especially on RAP-2.0 dataset. This finding indicates that the major issue with the external hardness prediction network is still the unbalanced distribution of attribute values and that DeepMAR weighted loss function is not completely capable of compensating it.

In conclusion, it can be stated that the realistic predictor approach using an external hardness predictor generally works. But the assumption that such an additional CNN is superior to the use of confidence scores cannot be fully validated for the pedestrian attribute task. Both networks learn complementary

Table 4.2: Realistic predictor results on RAP-2.0 dataset. Rejection strategies mainly improve instance metrics. Hardness scores provided by an explicit hardness predictor do not surpass the baseline given by using confidence scores of the classifier.

	RAP2.0			PA-100K		
Rejection Strategies	mA	F1	Rejected	mA	F1	Rejected
None	72.98	77.12	0.00	75.23	83.33	0.00
Hardness scores:						
Threshold	69.18	83.59	12.54	74.34	90.53	15.04
Quantile	66.07	81.80	24.58	74.20	88.08	22.54
mA	74.02	78.93	**7.75**	**78.09**	90.32	15.18
F1	66.14	79.52	16.68	74.78	87.67	**13.44**
Confidence scores:						
Threshold	71.77	**85.98**	14.51	75.87	**91.20**	17.32
mA	**74.79**	82.88	12.13	77.88	91.00	17.44

tasks and so the rejection rate is much lower when the hardness predictor network is used. However, results of the confidence score are not exceeded.

5 Conclusion and future work

This work aimed to apply the concept of realistic predictors to the field of pedestrian attribute recognition. The core idea was to address some of the biggest challenges in pedestrian attribute recognition while simultaneously achieving more reliable attribute estimates. To achieve this, the approach introduced in [17] was modified and optimized for the task of attribute recognition. This included, for instance, adapting the loss functions and alterations regarding to the network architecture. In addition, different strategies to determine meaningful thresholds for exclusion of unreliable predictions were proposed and extensively studied.

All in all the findings of this work showed that the concept of realistic predictors can be transferred to the field of pedestrian attribute recognition and accuracy improvements can be achieved. However, comprehensive experiments indicate that the predictions of hardness do not reflect the difficulty of the task equally well for all attributes. Especially attributes with strongly unbalanced value distributions in the training dataset cause problems and worsen the results. As a result, better performance was achieved if confidence scores are used instead of hardness predictions. In one point, however, the hardness predictions were strongly superior to the confidence values, namely in the number of rejected samples. From this it can be concluded that training a separate hardness predictor has its advantages.

In future research the training of the hardness predictor and the loss function can be improved in order to eliminate the imbalance problem of some attributes. The aim is to close the performance gap with the confidence-based rejection while maintaining the advantage in terms of number of rejected samples. Moreover, the hardness predictor approach allows to weight attributes during attribute-based person retrieval. By considering attributes according to their difficulty in predicting them during distance computation, incorrect retrieval results in early ranking positions can be avoided.

References

[1] Abrar H Abdulnabi et al. “Multi-task CNN model for attribute prediction”. In: *IEEE Transactions on Multimedia* 17.11 (2015), pp. 1949–1959.

[2] Esube Bekele and Wallace Lawson. “The Deeper, the Better: Analysis of Person Attributes Recognition”. In: *arXiv preprint arXiv:1901.03756* (2019).

[3] Ali Diba et al. “Deepcamp: Deep convolutional action & attribute mid-level patterns”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 3557–3565.

[4] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. “Actions and attributes from wholes and parts”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 2470–2478.

[5] Hao Guo, Xiaochuan Fan, and Song Wang. “Human attribute recognition by refining attention heat map”. In: *Pattern Recognition Letters* 94 (2017), pp. 38–45.

[6] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[7] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. “Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios”. In: *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. IEEE. 2015, pp. 111–115.

[8] Dangwei Li et al. “A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios”. In: *IEEE transactions on image processing* 28.4 (2018), pp. 1575–1590.

[9] Dangwei Li et al. “Pose guided deep model for pedestrian attribute recognition in surveillance scenarios”. In: *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2018, pp. 1–6.

[10] Yining Li et al. “Human attribute recognition by deep hierarchical contexts”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 684–700.

[11] Pengze Liu et al. “Localization guided learning for pedestrian attribute recognition”. In: *arXiv preprint arXiv:1808.09102* (2018).

[12] Xihui Liu et al. “Hydraplus-net: Attentive deep features for pedestrian analysis”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 350–359.

[13] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. “Deep imbalanced attribute classification using visual attention aggregation”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 680–697.

[14] M Saquib Sarfraz et al. “Deep view-sensitive pedestrian attribute inference in an end-to-end model”. In: *arXiv preprint arXiv:1707.06089* (2017).

[15] Arne Schumann, Andreas Specker, and Jürgen Beyerer. “Attribute-based Person Retrieval and Search in Video Sequences”. In: *Advanced Video and Signal Based Surveillance (AVSS), 2018 15th IEEE International Conference on*. 2018.

[16] Patrick Sudowe, Hannah Spitzer, and Bastian Leibe. “Person attribute recognition with a jointly-trained holistic cnn model”. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2015, pp. 87–95.

[17] Pei Wang and Nuno Vasconcelos. “Towards realistic predictors”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 36–51.

[18] Luwei Yang et al. “Attribute recognition from adaptive parts”. In: *arXiv preprint arXiv:1607.01437* (2016).