Pedestrian Attribute Classification in Surveillance: Database and Evaluation

Jianqing Zhu, Shengcai Liao, Zhen Lei, Dong Yi, Stan Z. Li*
Center for Biometrics and Security Research & National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences (CASIA)
95 Zhongguancun East Road, 100190, Beijing, China
{jqzhu, scliao, zlei, dyi, szli}@cbsr.ia.ac.cn

Abstract

Attributes are helpful to infer high-level semantic knowledge of pedestrians, thus improving the performance of pedestrian tracking, retrieval, re-identification, etc. However, current pedestrian databases are mainly for the pedestrian detection or tracking application, and semantic attribute annotations related to pedestrians are rarely provided. In this paper, we construct an Attributed Pedestrians in Surveillance (APiS) database with various scenes. The APiS 1.0 database includes 3661 images with 11 binary and 2 multi-class attribute annotations. Moreover, we develop an evaluation protocol for researchers to evaluate pedestrian attribute classification algorithms. With the APiS 1.0 database, we present two baseline methods, one for binary attribute classification and the other for multi-class attribute classification. For binary attribute classification, we train AdaBoost classifiers with color and texture features, while for multi-class attribute classification, we adopt a weighted K Nearest Neighbors (KNN) classifier with color features. Finally, we report and discuss the baseline performance on the APiS 1.0 database following the proposed evaluation protocol.

1. Introduction

In the past two decades, there is an explosive growth in surveillance video applications in the world. Surveillance cameras are installed in many places such as airports, train stations, parking lots, banks, etc. The videos captured by these cameras are very massive which presents formidable challenges for people to search useful information. Therefore, the smart video surveillance technology [21] has been used to process online video in real time for saving expensive human resources. Video analysis is a key module for smart video surveillance, including object detection [12, 25], object tracking [26], and object classification [13], etc.

*Corresponding author.

Figure 1. Thirty pedestrian examples in our database. Images in red rectangles are three examples of pedestrians with *female* and *short hair* attributes.

Recently, attributes (human-understandable properties, such as male, black eyes, long hair, etc.) are receiving more and more interests, since they are helpful to infer high-level semantic knowledge. Attributes based image representation was firstly proposed in [7] and has been successfully exploited in many recent works. For instance, Farhadi et al. [6] proposed an attribute-centric approach to detect unusual aspects of known objects and recognize unknown objects. Kumar et al. [14] used semantic attributes as mid-level features to aid face verification. In this application, the prediction model of the presence of each attribute on the input image is first learned, and the supervised object models on top of those attribute predictions are then built. Wang and Mori [24] proposed a method to jointly learn visual attributes, object classes and visual saliency in a unified framework, based on attribute interdependency. Meanwhile, attributes are also useful in zero-shot learning [15] and image retrieval [29].

In this work, we focus on the pedestrian attribute classification in video surveillance, where pedestrian attribute

classification can further provide useful information for applications such as pedestrian tracking, re-identification and retrieval. For example, Figure 1 shows that in the pedestrian retrieval application, the *female* and *short hair* attributes can effectively assist to find the desired pedestrians.

Pedestrian attributes in surveillance application include *gender*, *hair*, *clothing appearance*, *carrying thing*, etc. However, most of the existing pedestrian databases are in context of pedestrian detection or pedestrian tracking, which do not provide semantic pedestrian attribute annotations. Therefore, we construct an Attributed Pedestrians in Surveillance (APiS) database for pedestrian attribute classification research. Furthermore, we design a well established evaluation protocol for researchers to evaluate and compare their algorithms. At last, we provide baseline attribute classification methods and report their performance on the APiS 1.0 database following the evaluation protocol.

The rest of this paper is organized as follows. Section 2 summarizes some related works. Section 3 introduces the APiS 1.0 database and the evaluation protocol. Section 4 describes our baseline methods, including the binary attribute classification method and the multi-class attribute classification method. Section 5 shows experiments and analysis. Section 6 concludes this paper.

2. Overview of Related Work

Daniel et al. [22] proposed an attribute-based people searching system in surveillance environments. In this application, people are identified by a series of attribute detectors. The attributes used in this work include facial, hair, glass wearing, clothing color, etc. Each attribute detector is learned from large amounts of training data, however, their training data is not publicly available.

Layne et al. [16] utilized 15 semantic attributes as mid-level representations to aid person re-identification. The database used for attribute learning is a subset of the VIPeR [11] database. However, this subset only includes 632 annotated images captured in outdoor scene.

Bourdev et al. [2] proposed an approach to describe the appearance of people by using 9 binary attributes such as *male*, *T-shirt*, *long hair*, etc. They applied a subset of the Poselets [3] database for attribute learning. However, the database used in [2] includes many persons who are seating or standing in daily life, and many persons are only partially observable in the image. Therefore, this database is not specifically aimed at pedestrians in surveillance application.

Clothing appearance is an important attribute for people. Yang and Yu [28] proposed a clothing recognition system that identifies 8 clothing categories such as *suit* and *T-shirt* in a surveillance video, which was evaluated on a database including 937 persons and 25,441 cloth instances with 8 clothing categories. Chen et al. [4] proposed a method that comprehensively describes the upper clothing appearance with 23 binary attributes and 3 multi-class attributes. The clothing attribute database used in [4] includes 1,856 images. Liu et at. [19] collected a large online shopping database and a daily photo database for the research of the cross-scenario clothing retrieval. The online shopping database includes about 8,300 images and the daily photo database includes 4,300 images. The two database are labeled with 15 clothing attributes. However, for clothing description, the persons in [28, 4, 19] are mostly captured in frontal view. Furthermore, images in [28] were obtained by frontal face detection and tracking.

3. The APiS 1.0 Database

3.1. Database Composition

In order to construct a pedestrian database with comprehensive scenes, we select images from four sources: KITTI [10] database, CBCL Street Scenes [1] (CBCLSS for short) database, INRIA [5] database and SVS database (Surveillance Video Sequences at a train station collected by ourselves).

The KITTI database is captured by driving around in rural areas and on highways in Karlsruhe (a mid-size city in the state of Baden-Wrttemberg, southwest Germany), which has 14,999 images and the pixel resolution is 1242×375. The CBCLSS database is captured by DSC-F717 camera in the MA Street, Boston, which includes 3,547 images and the pixel resolution is 1280×960. The INRIA database is a famous database for pedestrian detection since 2005. The INRIA database has 902 images, and the pixel resolution of images in INRIA database varies from 176×257 to 1090×976.

The KITTI, CBCLSS and INRIA databases are captured in outdoor scene and their basic imaging angles are straight forward. In order to construct a more comprehensive database, we add an indoor scene database SVS extracted from four surveillance video sequences captured at a train station. Images in SVS database are captured by overhead-view surveillance cameras and their pixel resolution is 1920×1088.

Although the KITTI, CBCLSS and INRIA databases all provide the ground-truth bounding boxes of pedestrians, they follow different criterions to label the ground truth, which may cause each cropped pedestrian image including different proportion of background pixels. In order to reduce this adverse effect, a pedestrian detection approach [25] is performed to automatically locate candidate pedestrian regions. In this work, the width/height ratio of the pedestrian detector is set to be 0.376.

After pedestrian detection, we delete false positives and those too small images. That is, we only select images larger than 90 pixels in height and 35 pixels in width. We further resize all cropped pedestrian images into 128×48 pixels by

Table 1. The comparison between APiS 1.0 and other publicly available pedestrian attribute databases.

database	scene	for surveillance	#images	#attributes
APiS 1.0	outdoor indoor	yes	3661	13
a subset of VIPeR [16]	outdoor	yes	632	15
clothing attribute database [4]	outdoor	no	1,856	26

Table 2. The statistics of APiS 1.0 database.

attribute	sample distribution
male (positive/negtive/ambiguous)	2465 / 1121 / 75
long hair (positive/negtive/ambiguous)	408 / 3192 / 61
shirt (positive/negtive/ambiguous)	498 / 3052 / 111
T-shirt (positive/negtive/ambiguous)	1753 / 1805 / 103
long pants (positive/negtive/ambiguous)	2912 / 734 / 15
M-S (Medium and Short) pants (positive/negtive/ambiguous)	473 / 3135 / 53
long jeans (positive/negtive/ambiguous)	887 / 1858 / 916
skirt (positive/negtive/ambiguous)	192 / 3417 / 52
back bag (positive/negtive/ambiguous)	292 / 3302 / 67
S-S (Single Shoulder) bag (positive/negtive/ambiguous)	671 / 2880 / 110
hand carrying (positive/negtive/ambiguous)	514 / 3096 / 51
upper-body clothing color (black/white/gray/red/green/ blue/ambiguous/occluded/undefined)	786 / 788 / 135 / 280 / 132 / 309 / 464 / 193 / 574
lower-body clothing color (black/white/khaki/gray/ blue/occluded/ambiguous/undefined)	1539 / 120 / 166 / 170 / 886 / 138 / 388 / 254

bilinear interpolation. As a result, the APiS 1.0 database includes 3,661 images in total. Table 1 shows the comparison between the APiS 1.0 database and other publicly available pedestrian attribute databases.

3.2. Data Annotation

We manually accomplish attribute annotation, producing 11 binary and 2 multi-class attribute annotations on each image in the APiS 1.0 database. In this work, the two multi-class attributes are *upper-body clothing color* and *lower-body clothing color* attributes. Table 2 shows the statistics of each attribute in the APiS 1.0 database. In order to improve the accuracy and efficiency of attribute annotation, we divide the attribute annotation work into annotation stage and validation stage.

For binary attributes, at the annotation stage, each attribute is labeled independently. At the validation stage, we use the relationship between attributes to check the annotations. For example, a pedestrian with *long hair* is very likely to be a *woman*; *long pants* and *short pants* are mutually exclusive. Obviously, using these relationships can avoid artificial errors during the independent annotation stage. Additionally, at the validation stage, we label controversial samples with *ambiguous* annotation.

For multi-class attributes, at the annotation stage, we label images with 12 classes (*black, white, khaki, gray, red, green, blue, yellow, purple, ambiguous, occluded, undefined*). Here, the *undefined* class is composed of the samples that are not included in the previous 11 cases. For the *upper-body clothing color* attribute, the *occluded* class is mainly composed of the samples that are occluded by back bag, while for the *lower-body clothing color* attribute, the *occluded* class is mainly composed of the samples that are occluded by hand carrying. At the validation stage, we first merge the class that has less than 100 samples into the *undefined* class. Then, we vote the manual annotations of each color. If an annotation has 3 or more agreements, it will be accepted as the ground truth. Otherwise, it will be accepted as an *ambiguous* annotation. Finally, we obtain 6 *defined* colors (*black, white, gray, red, green, blue*) for the *upper-body clothing color* attribute and 5 *defined* colors (*black, white, khaki, gray, blue*) for the *lower-body clothing color* attribute, respectively. Figure 2 shows some annotation examples in our APiS 1.0 database.

3.3. Evaluation Protocol

We evaluate the performance of each attribute classification with 5-fold cross-validation. That is, we provide a sample index to separate the APiS 1.0 database into 5 equal sized subsets, and then evaluate each attribute classification based on the same sample index. The 5 results from the 5 folds are further averaged to produce a single performance report.

In the evaluation of the binary attribute classification, samples with *ambiguous* annotations are excluded. Two performance measures, the recall rate and false positive rate are applied for evaluation. The recall rate means the fraction of the correctly detected positives out of the whole positive samples, and the false positive rate represents the fraction of the mis-classified negatives out of the whole negative samples. The Receiver Operating Characteristic (ROC) curve is also adopted to compare different algorithms. At various threshold settings, a ROC curve can be drawn by plotting the recall rate vs. the false positive rate. Note that, our evaluation is based on cross-validation, therefore, we report the performance with the average ROC curve. In order to make a more intuitive performance report, the Area Under the average ROC Curve (AUC) is also used for evaluation. The larger the AUC is, the better the classification performance will be.

In the evaluation of the multi-class attribute classifica-

Figure 2. Some attribute annotation examples in the APiS 1.0 database. Where *M-S pants* is the abbreviation of Medium and Short pants, and *S-S bag* is the abbreviation of Single Shoulder bag. Our database is challenging because it has a large variations of viewpoint, pose, illumination and scene. In addition, only the straps of *back bag* can be seen in some cases, which is a challenging condition. Moreover, partially occlusion makes the classification of hand carrying more difficult.

tion, samples with *ambiguous* or *occluded* annotations are excluded. In order to handle unseen colors beyond the training data, we design an open-set identification [20] experiment to evaluate the performance of the multi-class attribute classification. In our open-set identification experiment, the *defined* colors are used as the gallery classes, while the *undefined* samples are used as negative samples beyond the *defined* colors. Therefore, in the testing phase, the *undefined* samples should be rejected as not belonging to the gallery, so that we know they are with other colors beyond the *defined* colors. To evaluate the open-set identification performance, we adopt the detection & identification rate P_{di} and the false positive rate P_{fp} defined in [20]. Assume that $\mathcal{G}$ represents a gallery set, $\mathcal{Q}_{\mathcal{G}}$ and $\mathcal{Q}_{\mathcal{N}}$ represents two probe sets. While $\mathcal{Q}_{\mathcal{G}}$ consists of classes in the gallery set $\mathcal{G}$ but with different images, $\mathcal{Q}_{\mathcal{N}}$ contains classes that are not present in $\mathcal{G}$. Then, P_{di} and P_{fp} are formulated as

$$P_{di}(t) = \frac{|\{q|q \in \mathcal{Q}_{\mathcal{G}}, score(q) \geq t, \text{and } cid(q) = 1\}|}{|\mathcal{Q}_{\mathcal{G}}|}, \tag{1}$$

$$P_{fp}(t) = \frac{|\{q|q \in \mathcal{Q}_{\mathcal{N}}, \text{and } score(q) \geq t\}|}{|\mathcal{Q}_{\mathcal{N}}|}, \tag{2}$$

where $score(q)$ is the decision score function that decides whether q is a *defined* sample, t is the decision threshold, and $cid(q)$ is the classification indicator which is equals to 1 if and only if q is correctly classified.

We also use the average ROC curve to report the open-set identification performance as in [17], where the ROC curve represents detection & identification rate vs. false positive rate at various threshold settings. In order to show the difference in performance of each *defined* color, we further separate the query images into several subsets, with each subset containing a single *defined* color and the *undefined* color. Then the above evaluation protocol is applied to each subset to draw a performance curve with respect to the *defined* color of that subset.

The constructed APiS 1.0 database as well as the evaluation protocol will be available on the project website: `http://www.cbsr.ia.ac.cn/english/APiS-1.0-Database.html`.

4. Baseline Methods

The aforementioned attributes include binary and multi-class attributes, therefore, we design baseline classification methods applicable for binary attribute classification and multi-class attribute classification, respectively. Regarding binary attribute classification, we extract color and texture features based on sliding window and then use Gentle AdaBoost [9] algorithm to train classifiers. Regarding multi-class attribute classification, we extract color features and use a weighted K Nearest Neighbors (KNN) algorithm to accomplish the multi-class classification.

4.1. Binary Attribute Classification

4.1.1 Feature Extraction

We apply a sliding window strategy for feature extraction. Specifically, we generate 3,697 sub-windows, with different sizes (whose height varies from 12 to 126 and width from 12 to 48, resulting in 66 different sizes) and a sliding step of 6 pixels. From each sub-window, the color, MB-LBP and HOG features are extracted.

Color Feature We build a joint color histogram in the HSV color space for each sub-window as the color feature. After uniform quantization, the resulting color histogram has 8, 3 and 3 bins in the H, S, and V color channels, respectively. The color histogram of each sub-window is normalized to be unit length under the ℓ_1 norm.

MB-LBP Feature Three MB-LBP [18] descriptors with 3×3, 9×9 and 21×21 scales are adopted. To reduce the feature dimension and achieve the rotation invariant property, we use uniform and rotation invariant patterns for MB-LBP code mapping, resulting in 10 bins in each MB-LBP histogram. As a result, a 30-dimensional MB-LBP feature is obtained for each sub-window by concatenating three MB-LBP histograms with different scales. The same normalization is applied as with the color feature.

HOG Feature Histogram of Oriented Gradient (HOG) [5] feature is also a texture descriptor and it has been successfully used in pedestrian detection. In this work, each sub-window is equally divide into 2×2 sub-regions, and in each sub-region a histogram of oriented gradient is calculated with 9 orientation bins. The HOG feature associated with each sub-window is obtained by concatenating the above four histograms into a 36-dimensional vector. The same normalization is applied as with the color and MB-LBP features.

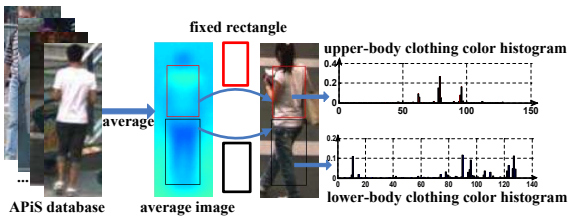

Figure 3. Feature extraction process for multi-class attribute classification. The red and black rectangle represent the region of upper body part and the region of lower body part, respectively.

4.1.2 AdaBoost Classifier Learning

After feature extraction, an image with 128×48 pixel resolution produces 266,184-dimensional color feature, 110,910-dimensional MB-LBP feature and 133,092-dimensional HOG feature. The feature dimension is very high, therefore, we need to reduce the feature dimension. We follow the work of [23], using the Adaboost [8] algorithm selects discriminative features to construct strong classifier one by one, which can accomplish feature selection and classifier learning at the same time.

In this work, we choose the Gentle AdaBoost [9] algorithm to accomplish feature selection because it usually has better performance than the discrete AdaBoost [8]. More specifically, we select the stump classifier with the minimum square error as the weak classifier in the Gentle AdaBoost algorithm. In order to take the advantage of multiple features, we also test the performance by concatenating the color, MB-LBP and HOG features at feature level fusion.

4.2. Multi-class Attribute Classification

4.2.1 Feature Extraction

For color attribute, in order to reduce the disturbance of background region, we need to determine the boundaries of upper body part and lower body part. Inspired by the method proposed in [27], we empirically label the regions of upper and lower body parts with 2 fixed rectangles on the average pedestrian image calculated from the whole database. In each rectangle region, a color histogram with 18, 5 and 5 bins in the H, S, and V color channels is extracted as the corresponding feature. The feature extraction process is illustrated in Figure 3. In our experiment, we set upper rectangle region [x, y, w, h]=[11, 20, 26, 42] and lower rectangle region [x, y, w, h]=[10, 64, 28, 54].

4.2.2 Weighted KNN Classifier Learning

We choose the weighted K Nearest Neighbors (KNN) algorithm as the baseline method for multi-class attribute classification. The weighted rule is following [30]. We further propose the decision confidence to make the weight-

ed KNN algorithm suitable for open-set identification problem.

Assume that $\mathbf{x}_0$ is a testing sample and y_0 is its class label to be predicted; $N = \{\mathbf{x}_k, k = 1, ..., \mathrm{K}\}$ represents the neighbors of $\mathbf{x}_0$; N_c and $\bar{N}_c$ represent the neighbors belonging to class c and not belonging to class c, respectively. In the open-set identification experiment, the decision score of the testing sample $\mathbf{x}_0$ is calculated according to Eq.(3). If $score(\mathbf{x}_0)$ is larger than a threshold t, $\mathbf{x}_0$ will be accepted as a gallery sample. Otherwise, it will be rejected. Then, if $\mathbf{x}_0$ has been identified as a gallery sample, it will be classified with Eq.(4). In our experiment, we set K to be 9.

$$score(\mathbf{x}_0) = \max_c(conf_c(\mathbf{x}_0)), \tag{3}$$

$$y_0 = \arg\max_c(conf_c(\mathbf{x}_0)), \tag{4}$$

where

$$conf_c(x_0) = W_{N_c} - W_{\bar{N}_c} = \sum_{\mathbf{x}_k \in N_c} w_k - \sum_{\mathbf{x}_k \in \bar{N}_c} w_k, \tag{5}$$

and

$$w_k = e^{-\|\mathbf{x}_k - \mathbf{x}_0\|}. \tag{6}$$

5. Experiments and Analysis

In the following, we evaluate the baseline performance of binary attribute classification and the multi-class attribute classification, respectively.

5.1. Performance of Binary Attribute Classification

Figure 4 shows the average ROC curve of each attribute. Color, MB-LBP, HOG and their fusion are tested and the corresponding AdaBoost classifier consists of 3000 weak classifiers. From Figure 4, we can find that color feature has better performance than texture feature on 6 clothing related attributes (*long jeans*, *long pants*, *M-S pants*, *shirt*, *skirt* and *T-shirt*), and HOG feature outperforms other two features on the rest attributes. Compared with the color feature, the HOG feature including pedestrian contour information is more informative for gender classification. For *long hair* attribute, the color histogram can not effectively reserve hair length information and its performance is not as good as texture features. Considering *back bag*, *S-S bag* and *hand carrying* attributes, color information is not robust enough to occlusion variations. As a result, the performance of texture features on these attributes outperform color features. Moreover, we can see that the fusion feature consistently outperforms three single types of feature on all the 11 binary attributes, which validates that the fusion of the color, MB-LBP and HOG features is useful for attribute classification.

Table 3 lists the average recall rate of binary attributes when the average false positive rate is 0.1. We can find that *long jeans*, *long pants* and *M-S pants* have good performance with average recall rate of 89.85%, 78.65% and 76.68%. These results indicate that pants related attributes in the lower body are easier to classify since they have fewer variations of position and shape appearance. In contrast, other attributes such as *back bag*, *S-S bag* and *hand carrying* with much more variations of viewpoint, position, size, occlusion and shape appearance are more difficult to classify. Figure 5 shows some samples that are difficult to be classified in the binary attribute classification when using AdaBoost classifiers and fusion features.

Table 3. Average recall rate of each binary attribute when using single type of feature and the fusion feature at the average false positive rate of 0.1.

attribute	recall rate(%)			
	color	MB-LBP	HOG	fusion feature
long jeans	88.95	54.45	55.47	**89.85**
M-S pants	71.46	54.12	53.28	**78.65**
long pants	70.71	60.06	56.87	**76.68**
skirt	58.33	51.56	60.42	**68.23**
male	41.95	34.12	46.69	**58.30**
back bag	39.38	43.15	47.60	**56.16**
T-shirt	44.72	36.05	38.22	**55.22**
long hair	40.20	34.56	49.51	**55.15**
shirt	47.39	38.96	42.97	**54.62**
hand carrying	43.58	35.99	44.16	**52.14**
S-S bag	30.70	37.85	36.96	**38.45**

Figure 5. Some samples that are difficult to be classified in the binary attribute classification, when using AdaBoost classifiers and fusion features. Many incorrectly classified *S-S bag* and *back bag* samples have only the sash parts appearing. Results are affected by similar clothing color appearing on shoulder region (*long hair*), occlusion (*hand carrying*) and weak illumination (*shirt* and *T-shirt*).

5.2. Performance of Multi-class Attribute Classification

The bottom right corner of Figure 4 shows the average ROC curves of the two multi-class attributes. We can see that *upper-body clothing color* and *lower-body clothing color* attributes have similar performance.

Figure 6 shows the performance of each color in *upper-body and lower-body clothing color* attributes. For the *upper-body clothing color* attribute, *black*, *red* and *white* have better performance, *blue* has medium performance, and *green* and *gray* have worse performance, because their AUC are decreasing. Among *blue* samples, dark blue sam-

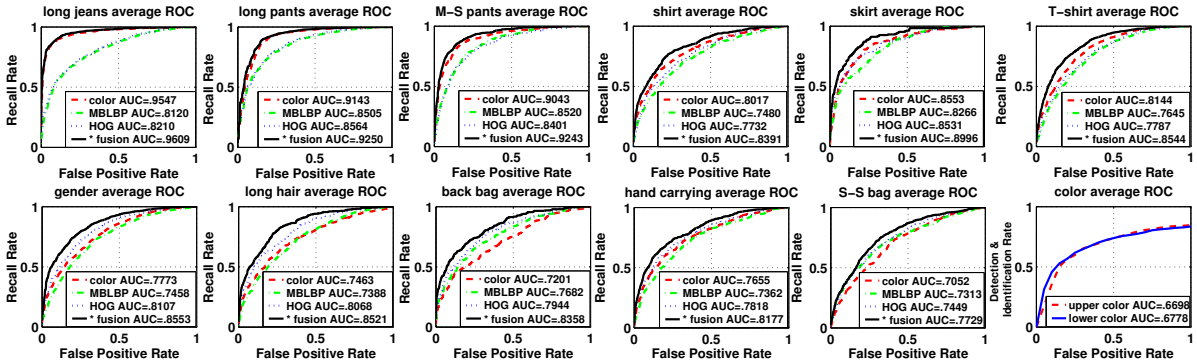

Figure 4. Average ROC curve of each attribute classification. Here, * indicates the corresponding average ROC curve has maximum AUC value.

ples are prone to be classified as *black* class, while light blue samples are easily to be assigned as *white* class. Regarding to the poor performances of the *green* and *gray* colors, the most possible reason is illumination variations. In weak illumination condition, *green* and *gray* samples are prone to be assigned as *black* class. In contrast, in strong illumination condition, they are tend to be classified as the *white* class.

Besides the same difficulties existing in the classification of *upper-body clothing color* attribute, the classification of *lower-body clothing color* attribute has two more difficulties: large background between legs and short clothing length. The background region between two legs of a striding pedestrian can disturb classification since the color histogram is extracted from a fixed rectangle region in the lower part of the pedestrian image. For example, when the background region appears to be *black*, then, a sample with the *gray lower-body clothing color* attribute is very likely to be mis-classified as the *black* class. Besides, using this color histogram extraction method, considering a pedestrian with short pants, the exposed skin region can also mislead the classification result. Figure 7 exhibits some samples that are difficult to be classified in the multi-class attribute classification.

In addition to the above reasons, there also exist two subjective factors which may cause poor classification performance. The first factor is the color perception difference of human eyes which causes ambiguous labeling of some confusing samples, and the other one is the number of color category that can be semantically defined is limited. These factors enlarge the intra-class difference which will cause the deterioration of color classification performance.

6. Conclusion and Future work

In this work, a attributed pedestrian database, namely APiS 1.0, which contains 3661 images with 13 semantic attribute annotations, has been collected and an evaluation protocol has been designed. Following the evaluation protocol, the AdaBoost classifier trained with color, MB-LBP and HOG features is evaluated to report the baseline performance of the binary attribute classification, and the weighted KNN classifier with color features is evaluated to report the baseline performance of the multi-class attribute classification. In the future, we plan to collect more images and increase the number of attribute annotation. We also try to use multi-label classification algorithms to take advantage of the relationships among attributes to improve the attribute classification performance. Finally, we hope that the APiS 1.0 database will promote the research of pedestrian attribute classification.

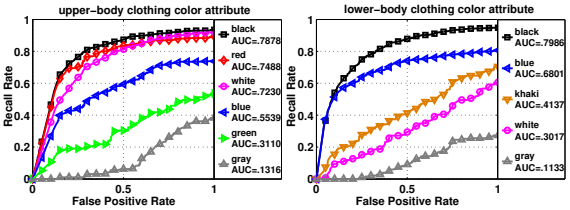

Figure 6. The curve of average recall rate and false positive rate for each color in *upper-body and lower-body clothing color* attributes.

Figure 7. Some samples that are difficult to be classified in multi-class attribute classification. Many mis-classified classified *blue* samples are dark blue and light blue samples. Results are affected by illumination condition (*green*, *gray*), large background between legs (*khaki*) and short clothing length (*white*).

7. Acknowledgement

This work was supported by the Chinese National Natural Science Foundation Project #61070146, #61105023,

#61103156, #61105037, #61203267, #61375037, National IoT R&D Project #2150510, National Science and Technology Support Program Project #2013BAK02B01, Chinese Academy of Sciences Project No. KGZD-EW-102-2, European Union FP7 Project #257289 (TABULA RASA), and AuthenMetric R&D Funds.

References

[1] S. M. Bileschi and L. Wolf. CBCL street scenes. 2006. http://cbcl.mit.edu/software-datasets/streetscenes.

[2] L. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. In *IEEE International Conference on Computer Vision*, pages 1543–1550, 2011.

[3] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *IEEE International Conference on Computer Vision*, pages 1365–1372, 2009.

[4] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *European Conference on Computer Vision*, pages 609–623. 2012.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*.

[6] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1778–1785, 2009.

[7] V. Ferrari and A. Zisserman. Learning visual attributes. In *Advances in Neural Information Processing Systems*, 2007.

[8] Y. Freund, R. E. Schapire, et al. Experiments with a new boosting algorithm. In *International Conf. on Machine Learning*, pages 148–156, 1996.

[9] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, 28(2):337–407, 2000.

[10] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *IEEE Conf. on Computer Vision and Pattern Recognition)*, pages 3354–3361, 2012.

[11] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE workshop on performance evaluation of tracking and surveillance*, 2007.

[12] C. Gu, P. Arbeláez, Y. Lin, K. Yu, and J. Malik. Multi-component models for object detection. In *European Conference on Computer Vision*, pages 445–458. 2012.

[13] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. In *European Conference on Computer Vision*, pages 459–472. Springer, 2012.

[14] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *IEEE Conf. on International Conference on Computer Vision*, pages 365–372, 2009.

[15] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 951–958, 2009.

[16] R. Layne, T. Hospedales, S. Gong, and Q. Mary. Person re-identification by attributes. In *British Machine Vision Conference*, 2012.

[17] S. Liao, A. K. Jain, and S. Z. Li. Partial face recognition: Alignment-free approach. In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pages 1193–1205, 2013.

[18] S. Liao, X. Zhu, Z. Lei, L. Zhang, and S. Z. Li. Learning multi-scale block local binary patterns for face recognition. In *Advances in Biometrics*, pages 828–837. Springer, 2007.

[19] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3330–3337, 2012.

[20] P. J. Phillips, P. Grother, and R. Micheals. Evaluation methods in face recognition. In *Handbook of Face Recognition*, pages 551–574. Springer, 2011.

[21] C. F. Shu, A. Hampapur, M. Lu, L. Brown, J. Connell, A. Senior, and Y. Tian. IBM smart surveillance system (s3): a open and extensible framework for event based surveillance. In *IEEE Conf. on Advanced Video and Signal Based Surveillance*, pages 318–323, 2005.

[22] D. Vaquero, R. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk. Attribute-based people search in surveillance environments. In *IEEE Workshop on Applications of Computer Vision*, 2009.

[23] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[24] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. In *European Conference on Computer Vision*, pages 155–168. 2010.

[25] J. Yan, Z. Lei, D. Yi, and S. Z. Li. Multi-pedestrian detection in crowded scenes: A global view. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3124–3129, 2012.

[26] B. Yang and R. Nevatia. Online learned discriminative part-based appearance models for multi-human tracking. In *European Conference on Computer Vision*, pages 484–498. Springer, 2012.

[27] H. Yang, S. Liao, L. Zhen, Y. Dong, and S. Z. Li. Exploring structural information and fusing multiple features for person re-identification. In *IEEE Workshop on Camera Networks and Wide Area Scene Analysis*, 2013.

[28] M. Yang and K. Yu. Real-time clothing recognition in surveillance videos. In *IEEE International Conference on Image Processing*, pages 2937–2940, 2011.

[29] F. X. Yu, R. Ji, M.-H. Tsai, G. Ye, and S.-F. Chang. Weak attributes for large-scale image retrieval. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2949–2956, 2012.

[30] J. Zavrel. An empirical re-examination of weighted voting for k-nn. In *Proceedings of the 7th Belgian-Dutch Conference on Machine Learning*, pages 139–148, 1997.