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.

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 al. [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-Württemberg, 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 [24] is performed to automatically locate candidate pedestrian regions. In this work, the width/height ratio of the pedestrian detector is set to 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.

<table>
<thead>
<tr>
<th>database</th>
<th>scene</th>
<th>for surveillance</th>
<th>#images</th>
<th>#attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>APiS 1.0</td>
<td>outdoor</td>
<td>yes</td>
<td>3661</td>
<td>13</td>
</tr>
<tr>
<td>a subset of VIPeR [16]</td>
<td>outdoor</td>
<td>yes</td>
<td>632</td>
<td>15</td>
</tr>
<tr>
<td>clothing attribute database [4]</td>
<td>outdoor</td>
<td>no</td>
<td>1,856</td>
<td>26</td>
</tr>
</tbody>
</table>

Table 2. The statistics of APiS 1.0 database.

<table>
<thead>
<tr>
<th>attribute</th>
<th>sample distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>2465 / 1121 / 75</td>
</tr>
<tr>
<td>long hair</td>
<td>408 / 3192 / 61</td>
</tr>
<tr>
<td>shirt</td>
<td>498 / 3052 / 111</td>
</tr>
<tr>
<td>t-shirt</td>
<td>1753 / 1805 / 103</td>
</tr>
<tr>
<td>long pants</td>
<td>2912 / 734 / 15</td>
</tr>
<tr>
<td>M-S (Medium and Short) pants</td>
<td>473 / 3135 / 53</td>
</tr>
<tr>
<td>long jeans</td>
<td>887 / 1858 / 916</td>
</tr>
<tr>
<td>skirt</td>
<td>192 / 3417 / 52</td>
</tr>
<tr>
<td>hand carrying</td>
<td>292 / 3302 / 67</td>
</tr>
<tr>
<td>S-S (Single Shoulder) bag</td>
<td>671 / 2880 / 110</td>
</tr>
<tr>
<td>upper-body clothing color (positive/ambiguous)</td>
<td>514 / 3096 / 51</td>
</tr>
<tr>
<td>lower-body clothing color (positive/ambiguous)</td>
<td>1539 / 120 / 166 / 170 / 886 / 138 / 388 / 254</td>
</tr>
</tbody>
</table>

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.

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 \( G \) represents a gallery set, \( Q_G \) and \( Q_N \) represents two probe sets. While \( Q_G \) consists of classes in the gallery set \( G \) but with different images, \( Q_N \) contains classes that are not present in \( G \). Then, \( P_{di} \) and \( P_{fp} \) are formulated as

\[
P_{di}(t) = \frac{|\{q| q \in Q_G, \text{score}(q) \geq t, \text{and} \ cid(q) = 1\}|}{|Q_G|},
\]

\[
P_{fp}(t) = \frac{|\{q| q \in Q_N, \text{and} \ \text{score}(q) \geq t\}|}{|Q_N|},
\]

where \( \text{score}(q) \) is the decision score function that decides whether \( q \) is a defined sample, \( t \) is the decision threshold, and \( \text{cid}(q) \) 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](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 $\ell_1$ norm.

MB-LBP Feature  Three MB-LBP [18] descriptors with $3 \times 3$, $9 \times 9$ and $21 \times 21$ 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 $2 \times 2$ 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.
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.

<table>
<thead>
<tr>
<th>attribute</th>
<th>color</th>
<th>MB-LBP</th>
<th>HOG</th>
<th>fusion feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>long jeans</td>
<td>68.95</td>
<td>54.43</td>
<td>53.47</td>
<td>89.85</td>
</tr>
<tr>
<td>M-S pants</td>
<td>71.46</td>
<td>54.12</td>
<td>53.28</td>
<td>78.65</td>
</tr>
<tr>
<td>long pants</td>
<td>70.71</td>
<td>60.06</td>
<td>56.87</td>
<td>76.68</td>
</tr>
<tr>
<td>skirt</td>
<td>58.33</td>
<td>51.56</td>
<td>60.42</td>
<td>68.23</td>
</tr>
<tr>
<td>male</td>
<td>41.95</td>
<td>34.12</td>
<td>46.69</td>
<td>58.30</td>
</tr>
<tr>
<td>back bag</td>
<td>39.38</td>
<td>43.15</td>
<td>47.60</td>
<td>56.16</td>
</tr>
<tr>
<td>T-shirt</td>
<td>44.72</td>
<td>36.05</td>
<td>38.22</td>
<td>55.22</td>
</tr>
<tr>
<td>long hair</td>
<td>40.20</td>
<td>34.56</td>
<td>49.51</td>
<td>55.15</td>
</tr>
<tr>
<td>shirt</td>
<td>47.39</td>
<td>38.96</td>
<td>42.97</td>
<td>54.62</td>
</tr>
<tr>
<td>hand carrying</td>
<td>43.58</td>
<td>35.99</td>
<td>44.16</td>
<td>52.14</td>
</tr>
<tr>
<td>S-S bag</td>
<td>30.70</td>
<td>37.85</td>
<td>36.96</td>
<td>38.45</td>
</tr>
</tbody>
</table>

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. 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.

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-
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.

7. Acknowledgement

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


