Learning to Recognize Pedestrian Attribute
Yubin Deng, Ping Luo, Chen Change Loy, Member, IEEE, and Xiaou Tang, Fellow, IEEE

Abstract—Learning to recognize pedestrian attributes at far distance is a novel research topic in video surveillance scenarios where face and body close-shots are hardly available; instead, only far-view video frames of pedestrian are given. In this study, we present an alternative approach that exploits the context of neighboring pedestrian images for improved attribute inference compared to the conventional SVM-based method. In addition, we conduct extensive experiments to evaluate the informativeness of background and foreground features for attribute recognition. Experiments is based on our newly released pedestrian attribute dataset, which is by far the largest and most diverse of its kind.

Index Terms—Large-scale database; attribute classification.

I. INTRODUCTION

Learning to recognize pedestrian attributes, such as gender, age, clothing style, has received growing attention in computer vision research, due to its high application potential in areas such as video-based business intelligence [13] and visual surveillance [4]. Real-world surveillance problems have presented themselves in scenarios where clear close-shots of face and body regions are seldom available in video frames. Thus, attribute recognition has to be performed at far distance using full body appearance (which can be partially occluded) in the absence of critical face/close-shot body visual information.

There are two fundamental challenges in attribute inference at far distance: 1) Appearance diversity - owing to diverse appearances of pedestrian clothing and uncontrollable multi-factor variations such as illumination and camera viewing angle, there exist large intra-class variations among different images for the same attribute. Learning to detect such attributes requires a rich set of training samples. Relying on a single source and small-scale training data would easily lead to an unrealistic model that generalizes poorly to unknown domains due to the inherent data bias. 2) Appearance ambiguity - far-view attribute recognition is an exceptionally difficult task due to inherent visual ambiguity and poor quality of visual features obtained from far view field (Fig. 1). In particular, an individual image may only occupy a few tens of imagery pixels whilst only a tiny fraction of them are truly distinctive for attribute classification. Often, parts of the body are occluded, either by obstacles or other pedestrians, which further increases the difficulty of extracting relevant features for inference. For instance, images with the ‘carrying backpack’ attribute may not necessarily have the full bag visible due to pedestrian posture (Fig. 1).

Existing datasets do not reflect the diversity nature in real-world environment. In view of this shortcoming, we introduced a new large-scale PEdesTrian Attribute (PETA) dataset1 in our previous work [3]. This dataset is by far the largest of its kind, covering more than 60 attributes on 19000 images. As can be seen from Fig. 1, in comparison with other datasets, PETA is more diverse and challenging in terms of imagery variations and complexity. More details are presented in Sec. II. Apart from further introducing the new dataset, we also investigate the informativeness of background and foreground features for recognition. Specifically, we follow the approach as in [3] and consider inference with the help from neighboring pedestrian images whose appearances look alike. We hypothesize that neighboring samples share natural invariance in their feature space, which could be treated as a form of regularization or context. As such, attribute inference of an image can be locally constrained by its neighbors to obtain a more reliable prediction. To this end, we view multiple pedestrian images as forming a Markov Random Field (MRF) graph. The underlying graph topology is automatically inferred, with node associations weighted by pairwise image similarity. The similarity can be estimated as the conventional Euclidean distance or more elaborated decision forest-based similarity with feature selection [17], [18]. By carrying out inference on the graph, we jointly reason and estimate the attribute probability of all images in the graph.

It is worth noting that MRF inference for smoothing [7] is commonly applied in image segmentation [12]. Inspired by the result in our previous work [3], we continue to explore this approach for pedestrian attribute inference. We summarize our contributions in this article as follows: 1) we further introduce the largest pedestrian attribute dataset to date for future research on attribute classification at far distance; 2)

Y. Deng, P. Luo, C. C. Loy and X. Tang are with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong. (e-mail: danny.s.deng.ds@gmail.com; lp011@ie.cuhk.edu.hk; ccloy@ie.cuhk.edu.hk; xtang@ie.cuhk.edu.hk)

1http://mmlab.ie.cuhk.edu.hk/projects/PETA.html
we provide more details on the benchmark performance by SVM-based method [6] and the MRF-based method [3]; and most importantly, 3) we evaluate the effect of pedestrian parsing and compare the informativeness of parsed regions. Thanks to features from different parsed regions as well as the exploited neighboring context, the performance of this new model surpass [3] by more than 4%. This fact shows that our new model has largely enhanced its capability to detect subtle attributes with satisfactory accuracy.

II. PEDESTRIAN ATTRIBUTE DATASET

A. Statistics and Uniqueness

The new PEdestriAn TriAttr (PETA) dataset\(^2\)\(^3\) comprises of 10 publicly available small-scale datasets. Specifically, it consists of 19000 images, with resolution ranging from 17 × 39 to 169 × 365 pixels. With erroneous images and duplicated copies removed, each image in PETA is newly labeled with 61 binary and 4 multi-class attributes. The binary attributes cover an exhaustive set of characteristics of interest, including demographics (e.g. gender and age range), appearance (e.g. hair style), upper and lower body clothing style (e.g. casual or formal), and accessories. The four multi-class attributes encompass 11 basic color namings [14], respectively, for footwear, hair, upper-body clothing, and lower-body clothing. The distribution of a binary attribute is considered balanced if the ratio of larger-to-smaller class is no more than 20:1. As such, out of the 61 binary attributes, 31 are balanced. Fig. 2(b) depicts the distribution of a few attributes with sample images\(^4\).

Compared to existing pedestrian attribute datasets, this new attribute dataset has three notable uniqueness: (1) Larger size: The size of the PETA dataset is over 5× and 15× larger than the APIS and VIPeR datasets, respectively. (2) High diversity: The composition of the new attribute dataset is enriched by smaller-scale datasets collected under different conditions from diverse scenes. As can be seen from Fig. 1 and summarized in the table in Fig. 2(a), despite that the constituents of PETA are all captured from far view field, they exhibit large differences in terms of lighting condition, camera viewing angles, image resolutions, background complexity, and indoor/outdoor environments. (3) Rich annotations: The PETA dataset contains far richer annotations in comparison with existing datasets, such as VIPeR [5], with only 15 binary attributes, and APIS [16], with 11 binary and 2 multi-class attributes. It is worth pointing out that the 61 annotated attributes in PETA dataset include the 15 attributes that are suggested by the UK Home Office and UK police to be the most valuable in tracking and criminal identification [11].

B. Usage

This dataset can serve as an alternative benchmark and it can be used in visual surveillance research on pedestrian tracking, detection, re-identification, and activity analysis. In the next section, we present two benchmarking methods for attribute classification, a fundamental task in visual understanding.

III. BASELINE METHODS

A. Baseline 1

SVM with intersection kernel (iKVM) [10]\(^5\) reduces both the time and space complexities from \(O(nm)\) of the traditional linear kernel SVM to \(O(n)\). Here, \(m\) and \(n\) are the number of the support vectors and the dimension of feature vectors. Previous study [6] has applied this method successfully for pedestrian attribute classification. Cross validation for slack parameter \(C\) is performed as in [6].

B. Baseline 2

To improve attribute inference, we exploit the context of neighboring images by Markov Random Field (MRF), which is an undirect graph, where each node represents a random variable and each edge represents the relation between two connected nodes. The energy function of MRF over a graph \(G\) can be defined as follows

\[
E_{MRF}(G) = \sum_{u \in G} C_u(l_u) + \sum_{u \in G} \sum_{v \in N(u)} S_{uv}(l_u, l_v),
\]

where \(u, v \in G\) are two random variables in the graph and \(l_u\) denotes the state of \(u\). \(C_u\) and \(S_{uv}\) signify the unary cost and pairwise cost functions, respectively. More precisely, they indicate the cost of assigning state \(l_u\) to variable \(u\) as well as the cost of assigning states to neighboring nodes \(u, v\), which is determined based on the graph structure (e.g., assigning different states to nodes that are similar is penalized). \(N(u)\) is a set of variables that are the neighbors of \(u\).

\(^2\)Images in PETA dataset are all exclusive from those in APIS [16].

\(^3\)All images in PETA are freely available for academic use except i-LIDS, which requires application to United Kingdom Home Office, https://www.gov.uk/imagery-library-for-intelligent-detection-systems.

\(^4\)Sample images of the datasets and the full distributions of all attributes can be found in http://mmlab.ie.cuhk.edu.hk/projects/PETA.html.

\(^5\)http://www.cs.berkeley.edu/~smaji/projects/fiksvm/
In this work, each random variable corresponds to an image and the relation between two variables corresponds to the similarity between images. The states of variable are the values of the image attribute, which is \( l_u \in \{0, 1\} \). The unary function is modeled by
\[
C_u(l_u) = -\log P(l_u|u),
\]
where \( P(l_u|u) \) is the probability of predicting the attribute value of image \( u \) as \( l_u \). This probability is learned by ikSVM.

Now we consider the definition of the pairwise function. To define affinity between nodes, a simple way widely adopted by existing methods, such as [15], is the Gaussian kernel, \( \exp(-\frac{|u-v|^2}{\sigma^2}) \), in which \( u, v \) indicate the feature vectors of two images and \( \sigma \) is a coefficient that needs to be tuned. The graph built on this kernel function can model the global smoothness among images. However, when large variations are presented, one may consider modeling the local smoothness and discovering the intrinsic manifold of the data. Thus, an alternative is to employ the random forest (RF) [2] to learn the pairwise function [17], [18]. The RF we adopted is unsupervised, i.e., it takes unlabeled test samples as input. The output is pairwise sample similarity derived from the data partitioning discovered at the leaf nodes of RF. The unsupervised RF can be learned using the pseudo two-class method as in [17], [18], [8]. Specifically, we treat the original unlabeled test samples as first class. The pseudo second class is created by sampling at random from the univariate distributions of the unlabeled test samples. With this strategy, the unsupervised RF learning problem becomes a canonical classification problem that can be solved by conventional classification forest training method. Specifically, the information gain of unsupervised RF is identical to that of conventional supervised RF, defined as
\[
\Delta I = I_p - \frac{n_l}{n_p} I_l - \frac{n_r}{n_p} I_r,
\]
where \( p, l, \) and \( r \) refer to a splitting node and its left and right child. The variable \( n \) denotes the number of samples at a node, \( n_p = n_l + n_r \). \( I \) is the Gini impurity measure that is used for deciding the ordering of features at each node [2]. As such, the Gini measure is extended to cope with the unsupervised data.

The pairwise function is expressed as
\[
S_{uv}(l_u, l_v) = \begin{cases} \frac{1}{T} \sum_{t=1}^{T} \exp(-\text{dist}^t(u,v)) & \text{if } l_u \neq l_v, \\ 0 & \text{otherwise.} \end{cases}
\]
Here, \( \text{dist}^t(u,v) = 0 \) if \( u, v \) fall into the same leaf node and \( \text{dist}^t(u,v) = +\infty \) otherwise, where \( t \) is the index of tree. Since the graph is dense, the inference of MRF is difficult. Thus, we build a \( k \)-NN sparse graph by limiting the number of neighbors for each node. We set \( k = 5 \) in our experiment. Eq.(1) can be efficiently solved by the min-cut/max-flow algorithm introduced in [1].

IV. EXPERIMENTS

We present benchmark results on PETA by evaluating the performance of intersection kernel SVM (ikSVM) [10], MRF with Gaussian kernel (MRFg), and MRF with random forest (MRFr), as discussed in Sec.III.

We randomly partitioned the dataset images into 9,500 for training, 1,900 for verification and 7,600 for testing. We selected 35 attributes for our study, consisting of the 15 most important attributes in video surveillance proposed by human experts [6], [11] and 20 difficult yet interesting attributes. For the attributes ‘glasses’ and ‘v-neck’ have a limited number of positive examples. For the attributes with unbalanced positives and negatives samples, we trained ikSVM for each attribute by augmenting the positive training examples to the same size as negative examples with small variations in scale and orientation. This is to avoid bias due to imbalanced data. For MRFg and MRFr, we built the graphs using two different schemes. The first scheme, symbolized by MRFg1 and MRFr1, is to construct the graphs with only the testing images. The second one, symbolized by MRFg2 and MRFr2, is to include both training and testing samples in the graphs.

A. Features

Low-level color and texture features have been proven robust in describing pedestrian images [6], including 8 color channels such as RGB, HSV, and YCbCr, and 21 texture channels obtained by the Gabor and Schmid filters on the luminance channel. The setting of the parameters of the Gabor and Schmid filters are given in [6]. We horizontally partitioned the image into six strips and then extracted the above feature channels, each of which is described by a bin-size of 16.

B. Results

Feature extraction with pedestrian parsing: It is intuitive to recognize the desired attributes from the pedestrian foreground regions. Would background regions play any role? We wish to examine if discarding background region would facilitate more accurate pedestrian attribute recognition. To this end, we apply the Deep Decompositional Network [9] to parse a pedestrian image into different body regions. Regions such as hair, face, body, arms, and legs of the pedestrian constitute the foreground regions and the remaining regions are considered
as the background. Some examples of foreground/background are depicted in Fig. 3.

Given the extracted foreground and background, we compare different feature extraction schemes for appearance representation: (1) features from the whole image space, (2) features from the foreground region only, (3) concatenating features from the foreground and background regions, and (4) concatenating features from the foreground region and whole image space. These schemes are denoted by ‘whole’, ‘fore’, ‘fore+back’, ‘fore+whole’, respectively.

As shown in Table I, we observe that simply extracting the foreground features (‘fore’) results in an inferior performance than that resulted from using the whole image. It suggests that background information is critical in facilitating the detection of attributes. If we inspect the classification results of each attribute in detail, we observe that background plays a pivotal role for recognizing ‘Backpack’, ‘CarryingOther’, ‘Plastic bag’, and ‘No carrying’ attributes. This is reasonable since the visual evidence that corresponds to these attributes is not only captured by the pedestrian foreground region. Moreover, slight drops of accuracy are observed on cloth-style related attributes, e.g. ‘Jeans’ and ‘Trousers’, if features are only extracted from the foreground. These results all suggest that background region could provide context for better attribute recognition performance.

Extracting and concatenating features from the foreground and background (‘fore+back’) sees a slight improvement for easy-to-spot attributes such as ‘AgeAbove60’, ‘Casual upper wear’, ‘Formal upper wear’, ‘Hat’, ‘Jeans’, ‘Long hair’, ‘Male’, ‘Shorts’, ‘Skirt’; however, the performance deteriorates for other attributes due to the inevitable noise contained in the features extracted solely from the background. Finally, when ‘fore+whole’ is adopted, a significant boost in the performance is observed, even for hard-to-spot attributes like ‘Leather Shoes’ and ‘Plastic bag’. The ‘fore+whole’ scheme seems to better exploit the information provided by the background.

Evaluating the importance of neighboring context: We choose the best three of the four feature extraction schemes, namely the ‘whole’, ‘fore+back’, and ‘fore+whole’, and evaluate our proposed method for detecting pedestrian attributes. We report the attribute detection accuracy in Table II and list some further observations as follows.

Firstly, the MRF-based methods outperform ikSVM on most of the attributes. For instance, MRFR2 achieves an average of 3.4% improvement over ikSVM for the ‘age’ attributes shown on the top of the table. This is significant in a dataset with large appearance diversity and ambiguity and it demonstrates that graph regularization can improve attribute inference. Plus, an about 5% boost of performance is observed for attributes such as ‘MessengerBag’, ‘No accessory’, ‘No carrying’, and ‘Trousers’ and we observe a near 10% boost over ikSVM for ‘carryingOther’ and ‘Shoes’. Secondly, the MRF graphs built with the second scheme (graph is constructed by both train and test samples) is superior compared to the first scheme (graph is constructed by test samples only). This is reasonable because using both the training and testing data can better cover the image space. Third, for many important attributes, such as ‘Trousers’ and ‘Shoes’, random forest works much better than Gaussian kernel.

Moreover, we observed that for our proposed MRF methods, the importance of background information as context is best exploited when the inherent background noise is reduced. This observation corresponds with the detection performance using ikSVMs and we show that the best result is obtained when we concatenate features extracted separately on foreground and the whole image (‘fore+whole’), which on average outperforms the ‘whole’ scheme by 4.4%. Still, the detection performances for ‘sunglasses’ and ‘stripes’ have dropped mainly due to the large noise outside the corresponding attribute region as well as insufficient positive samples for model training.

Fig. 4 shows some attribute classification results using the forest MRF. The detection performance is satisfactory for most attributes with complement information from the neighboring context. False negative samples typically result from occlusion (e.g. backpack), color ambiguity (long hair)

| Attribute       | whole | fore | fore+back | fore+whole |
|-----------------|-------|------|-----------|------------|
| Age16-30        | 80.4  | 78.6 | 78.9      | 83.1       |
| Age31-45        | 73.6  | 71.9 | 71.8      | 77.6       |
| Age46-60        | 73.1  | 72.6 | 72.3      | 79.1       |
| AgeAbove60      | 87.2  | 89.5 | 89.1      | 93.5       |
| Backpack        | 66.7  | 64.4 | 65.6      | 70.7       |
| CarryingOther   | 64.6  | 59.4 | 59.7      | 66.9       |
| Casual lower    | 70.7  | 69.4 | 70.1      | 76.5       |
| Casual upper    | 70.3  | 69.7 | 71.0      | 76.0       |
| Formal lower    | 71.0  | 69.1 | 70.4      | 76.6       |
| Formal upper    | 70.0  | 69.2 | 70.3      | 76.8       |
| Hat             | 82.3  | 81.6 | 83.2      | 89.4       |
| Jacket          | 67.7  | 63.9 | 64.7      | 69.6       |
| Jeans           | 74.9  | 74.1 | 75.8      | 79.8       |
| Leather Shoes   | 78.9  | 76.9 | 77.9      | 84.0       |
| Logo            | 51.1  | 50.0 | 50.3      | 53.4       |
| Long hair       | 71.5  | 73.6 | 74.5      | 79.4       |
| Male            | 79.7  | 80.3 | 80.0      | 84.6       |
| MessengerBag    | 71.8  | 68.9 | 68.8      | 74.8       |
| Muffler         | 88.0  | 85.9 | 86.9      | 92.2       |
| No accessory    | 76.8  | 74.1 | 74.0      | 79.2       |
| No carrying     | 70.4  | 66.7 | 68.0      | 72.5       |
| Plaid           | 64.0  | 60.6 | 59.3      | 65.1       |
| Plastic bag     | 74.9  | 71.9 | 73.2      | 79.0       |
| Sandals         | 50.6  | 50.6 | 51.3      | 51.9       |
| Shoes           | 70.6  | 67.1 | 66.5      | 72.0       |
| Shorts          | 56.0  | 60.5 | 61.6      | 65.2       |
| ShortSleeve     | 71.3  | 68.4 | 69.9      | 75.1       |
| Skirt           | 64.0  | 65.9 | 64.7      | 69.6       |
| Sneaker         | 67.5  | 64.9 | 65.5      | 71.5       |
| Stripes         | 51.5  | 50.4 | 50.0      | 51.9       |
| Sunglasses      | 53.2  | 51.3 | 52.6      | 53.3       |
| Trousers        | 74.0  | 73.2 | 73.0      | 77.9       |
| Tshirt          | 64.3  | 61.0 | 62.8      | 71.1       |
| UpperOther      | 80.7  | 79.0 | 79.4      | 83.2       |
| V-Neck          | 51.1  | 51.7 | 53.3      | 53.3       |
| **AVERAGE**     | 69.5  | 68.2 | 68.8      | 73.6       |
Table 2: Classification accuracy of different methods.

| Attribute         | MRFg1 | MRFg2 | MRFr1 | MRFr2 |
|-------------------|-------|-------|-------|-------|
| Age 16-30         | 80.9  | 78.9  | 81.7  | 80.9  |
| Age 31-45         | 74.6  | 72.3  | 76.2  | 74.0  |
| Age 46-60         | 71.4  | 72.6  | 75.2  | 73.2  |
| Age Above 60      | 87.2  | 89.2  | 88.2  | 86.3  |
| Backpack           | 67.1  | 65.9  | 67.1  | 67.0  |
| Carrying Other    | 64.8  | 60.2  | 66.8  | 64.6  |
| Casual lower      | 70.9  | 69.8  | 71.6  | 70.4  |
| Casual upper      | 70.4  | 70.7  | 71.3  | 69.8  |
| Formal lower      | 71.2  | 70.9  | 71.8  | 70.5  |
| Formal upper      | 70.3  | 70.7  | 70.4  | 70.3  |
| Hat               | 82.9  | 83.2  | 84.3  | 82.3  |
| Jacket            | 68.3  | 65.0  | 68.4  | 68.1  |
| Jeans             | 75.2  | 76.3  | 76.1  | 75.0  |
| Leather Shoes     | 80.1  | 78.0  | 80.9  | 79.1  |
| Logo              | 51.1  | 50.5  | 51.0  | 51.1  |
| Long hair         | 71.7  | 75.2  | 72.6  | 71.8  |
| Male              | 80.3  | 79.9  | 80.9  | 80.6  |
| Messenger Bag     | 72.9  | 69.0  | 74.3  | 72.7  |
| Muffler           | 88.3  | 86.9  | 89.5  | 86.5  |
| No accessory      | 77.2  | 73.8  | 78.6  | 77.1  |
| No carrying       | 70.6  | 68.5  | 71.6  | 70.6  |
| Pail              | 64.5  | 59.6  | 64.5  | 65.0  |
| Plastic bag       | 74.9  | 73.6  | 75.5  | 73.9  |
| Sandals           | 50.8  | 51.2  | 50.6  | 50.6  |
| Shoes             | 71.0  | 66.9  | 72.5  | 70.8  |
| Shorts            | 56.5  | 61.8  | 56.5  | 56.5  |
| Short Sleeve      | 71.7  | 70.5  | 71.8  | 71.8  |
| Skirt             | 64.0  | 65.3  | 64.0  | 64.0  |
| Steaker           | 68.1  | 66.2  | 69.0  | 68.2  |
| Stripes           | 52.3  | 50.0  | 52.3  | 52.3  |
| Sunglasses        | 53.2  | 52.6  | 53.2  | 53.9  |
| Trouser           | 74.5  | 72.9  | 75.7  | 75.7  |
| T-shirt           | 64.5  | 63.6  | 64.6  | 63.6  |
| Upper Other       | 80.7  | 79.3  | 81.8  | 81.1  |
| V-Neck            | 51.1  | 53.3  | 51.1  | 51.1  |
| Average           | 69.9  | 69.0  | 70.6  | 69.7  |

There are three small columns for each compared methods. They correspond to the three feature extraction schemes, i.e., 'whole', 'forehead-back', and 'forehead' respectively.

References:
[1] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. TPAMI, 2004.
[2] L. Breiman. Random forests. Machine learning, 2001.
[3] Y. Deng, P. Luo, C. C. Loy, and X. Tang. Pedestrian attribute recognition at far distance. ACM MM, 2014.
[4] S. Gong, M. Cristani, C. C. Loy, and T. Hospedales. The re-identification challenge. In S. Gong, M. Cristani, S. Yan, and C. C. Loy, editors, Person Re-Identification. Springer, 2013.
[5] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In IEEE PETS, 2007.
[6] R. Layne, T. M. Hospedales, S. Gong, et al. Person re-identification by attributes. BMVC, 2012.
[7] Z. Li. Markov random field modeling in computer vision. Springer-Verlag New York, Inc., 1995.
[8] B. Liu, Y. Xia, and P. S. Yu. Clustering through decision tree construction. In CIKM. ACM, 2000.
[9] P. Luo, X. Wang, and X. Tang. Pedestrian parsing via deep decompositional neural network. ICCV, 2013.
[10] S. Maji, A. C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. CVPR, 2008.
[11] T. Nortcliff. People analysis cctv investigator handbook. Home Office Centre of Applied Science and Technology, 2011.
[12] C. Rother, T. Minka, A. Blake, and V. Kolmogorov. Co-segmentation of image pairs by histogram matching-incorporating a global constraint into MRFs. CVPR, 2006.
[13] C. Shan, F. Porikli, T. Xiang, and S. Gong. Video Analytics for Business Intelligence, volume 409. Springer, 2012.
[14] J. Van De Weijer, C. Schmid, J. Verbeek, and D. Larlus. Learning color names for real-world applications. TIP, 18(7):1512–1523, 2009.
[15] Z.-J. Zha, X.-S. Hua, T. Mei, J. Wang, G.-J. Qi, and Z. Wang. Joint multi-label multi-instance learning for image classification. CVPR, 2008.
[16] J. Zhu, S. Liao, Z. Lei, D. Yi, and S. Z. Li. Pedestrian attribute classification in surveillance: Database and evaluation. In ICCV Workshop, 2013.
[17] X. Zhu, C. C. Loy, and S. Gong. Video synopsis by heterogeneous multi-source correlation. In ICCV, 2013.
[18] X. Zhu, C. C. Loy, and S. Gong. Constructing robust affinity graph for spectral clustering. In CVPR, 2014.

Fig. 4. Examples of attribute classification with forest MRF (MRFr2).