Add-on Strategies for Fine-grained Pedestrian Classification

Yoshihito Kokubo*, Yu Wang†, Jien Kato*, Guanwen Zhang‡ and Kenji Mase*
*Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, Aichi, Japan
Email: kokubo@cmc.ss.is.nagoya-u.ac.jp, ywang@mv.ss.is.nagoya-u.ac.jp
jien@is.nagoya-u.ac.jp, mase@nagoya-u.jp
†School of Electronics and Information, Northwestern Polytechnical University
Xi’an, Shanxi, China
Email: zhang@mv.ss.is.nagoya-u.ac.jp

Abstract—In this paper, we present four add-on strategies for the fine-grained pedestrian classification task. These strategies are: (1) super-resolution based image preprocessing, which helps to recover the image details; (2) patch dividing based deep feature extraction, which extracts features in a way that preserves the spatial layout of input images; (3) pose-wise classifier sharing, which learns robust classifiers and makes robust predictions using pose information; and (4) graphical model based inference, which utilizes the interdependence between different subcategories to update raw estimations. The proposed strategies are independent and flexible, which make it easy to implement them in practice. We evaluated these strategies on the CRP dataset [1] and confirmed that all of them lead to improvements over the baseline. We also confirmed an improvement over the state-of-the-art when all strategies are combined together.

I. INTRODUCTION

The objective of fine-grained pedestrian classification is to recognize fine-grained categories of pedestrians, which are also called subcategories, such like sex, age, weight, and so on. Such a technique leads to higher level understanding on pedestrians as well as their activities, therefore is a perspective key component in the next generation smart vehicles and surveillance systems. Because of its importance, in recent years, more and more research efforts have been devoted to this task in the computer vision community [2] [1].

Even in case that classification subject is not pedestrian, fine-grained object classification itself is a very challenging task. Different with generic object classification, in which categories are visually distinct, the subcategories in fine-grained classification are usually very similar in their visual appearance. As a result, using a large number of various image statistics can not always lead to good performance. How to effectively capture subtle differences in the classification process becomes crucial for achieving high classification performance. Thanks to the recent advances in deep learning [3] [4], we have observed significant improvements of fine-grained classification on object categories such like dogs [5], birds [6], flowers [7] and so on.

The task becomes more difficult when the subject becomes pedestrians. For pedestrians, additional difficulties are introduced by the low image quality, large variation of poses, exten-
Within the pipeline, super-resolution (SR) and patch dividing (PD) are utilized to improve the feature representation. SR recovers details from the inputs, while PD makes sure that the feature representations reflect the spatial layout of pedestrians. On the other hand, pose-wise classifier sharing (PS) and graphical model (GM) are used to improve the classification performance. PS improves the robustness of learned classifiers, while GM utilizes the interdependence between subcategories to improve the prediction accuracy. In practice, the above four strategies can be utilized either individually or combinationally, which lead to easy practical usages. In the follow, we will discuss them in detail.

A. Super Resolution and Patch Dividing

In general, fine-grained pedestrian classification algorithms take pedestrian detectors’ outputs as inputs [2] [1]. Since pedestrian detectors usually output pedestrian instances in different sizes and aspect ratios [8], it is necessary to perform normalization before conducting feature extraction. In the literature, the most popular normalization strategy is to resize the images to a fixed size by interpolation [9]. However, such a strategy does not only lead to blurred images, but also changes the inputs’ aspect ratio, which all may hamper the fine-grained classification performance.

In this work, we propose to utilize super resolution and patch dividing to avoid the above issues. For an input image \( I_1 \) of size \( h_1 \times w_1 \), we first conduct super-resolution to up-sample it to \( I_h \) of size \( \frac{h_1 w_1}{w_0} \times w_0 \), using fixed width \( w_0 \) and unchanged aspect ratio \( \frac{h_1}{w_1} \). We adopt the method in [10] for super-resolution, in which sparsity prior is utilized. Specifically, a high-resolution image patch is assumed to have a sparse representation, and this sparse representation is recovered from the low-resolution image patch’s sparse representation using a pair of co-trained low-resolution and high-resolution dictionaries.

In order to deal with different \( \frac{h_1 w_1}{w_0} \) of \( I_h \), we further divide \( I_h \) horizontally into three square patches: \( \text{Patch}_1 = I_h(1: w_h, :) \), \( \text{Patch}_2 = I_h(\frac{(h_h-w_h)w_0}{2w_1}, : \frac{(h_h+w_h)w_0}{2w_1}, :) \), and \( \text{Patch}_3 = I_h(\frac{(h_h-w_h)w_0}{w_0}, : \frac{h_h w_0}{w_0}, :) \). In this work, we use pre-trained convolutional neural networks (CNN) as the feature extractor for their high performance in image classification [11]. For each patch, we feed it to a pre-trained CNN and take the response after the first fully connected layer (FC6) and non-linear gating of the Rectified Linear Unit, which is a 4,096 dimensional non-negative vector. The CNN features of the three patches are further concentrated together to a 12,288 feature vector to represent the input image. Such a vector representation is computed from enhanced image with unchanged aspect ratio, which is supposed to serve better in the fine-grained classification task.

B. Pose-wise Classifier Sharing

Different with other object categories such like birds [6] and flowers [7], pedestrians have relative larger variation in their pose appearance. On the other hand, fine-grained pedestrian classification approaches usually involve classifier training for multiple detailed subcategories. Since for each subcategory, pedestrians of different class labels may appear to have the same pose, while pedestrians of the same class label may appear to have different poses, such conflicts may miss-guide the classifier training process and lead to classifiers that are sensitive to the pose variation.

The pose-wise classifier sharing (PS) strategy is developed to deal with this issue. The basic ideas underlying such a strategy are: (1) classifiers that learned from pedestrians of similar poses are more robust; (2) classification using classifiers learned from pedestrian instances of similar poses leads to better performance. With the PS strategy, for each subcategory, pose-wise classifiers are learned in the learning phase, and some of their predictions are shared to get the final classification outputs in the predicting phase.

Suppose the training set for one of the subcategory is \( D = \{ (f_i, p_i, c_i) \}_{i=1}^n \), where \( n \) is the number of training samples, \( f_i \) denotes the image feature, \( p_i \) specifies the pose configuration (\( p_i \) could be obtained by manual labeling operation [1] or using a pose estimation algorithm such like...
In order to avoid the final results over-dependent on the graphical model, we did not implement full inference over all possible combinations of \((i, j, k)\). If a classifier output is of sufficient confidence, we just trust it and only conduct inference for the remaining subcategories and classes. We use a cut-off threshold \(\delta\) to determine if the classification outputs

\[ p(C_i^a, C_j^w, C_k^w) = \frac{1}{Z} \varphi_{a,w}(C_i^a, C_j^w) \varphi_{w,z}(C_j^w, C_k^w) \]

\[ = 1 \sum_{x \in \{0,1\}} \psi_{x,y}(C_i^a | C_j^w) + \psi_{y,z}(C_j^w | C_k^w) \]

\[ \approx 1 \sum_{x \in \{0,1\}} \psi_{x,y}(C_i^a | C_j^w) + \psi_{y,z}(C_j^w | C_k^w) \]

\[ \psi_{x,y}(C_i^a | C_j^w) + \psi_{y,z}(C_j^w | C_k^w) \]

where \(\psi_{x,y}\) represents the correlation between subcategories \(x\) and \(y\). It is approximated using \(\psi_{x,y}\) and \(\psi_{y,z}\), which are learned from the training data by calculating the co-occurrences of their subcategory classes, \(p(C_i^a, C_j^w, C_k^w)\) is the probability of belonging to the class \(u\) in the subcategory \(x\). It can be obtained by mapping the classification output of the \(u\)th classifier of the subcategory \(x\) to the \([0,1]\) interval with linear regression. The inference process can be done by evaluating all combinations of \((i, j, k)\), and selecting the best set \((i^*, j^*, k^*)\) that has the maximum \(p(C_i^{a*}, C_j^{w*}, C_k^{w*})\).
of a specific subcategory worth trust. There are in total three such thresholds, which were learned from the training data by minimizing the training error through cross-validation.

III. Experiment Results

We evaluated the proposed add-on strategies on a large scale pedestrian dataset. We studied both the individual effectiveness and the combinational effectiveness of these strategies. The details are presented in the follow.

A. Dataset and Configurations

The experiments were implemented on the Caltech Roadside Pedestrians (CRP) dataset [1], which consists of 7 videos that are captured from a moving vehicle. The dataset contains 32,457 bounding box annotations of pedestrians. 27,454 of them were labelled with 14 body part keypoints and 4 fine-grained subcategories (sex: female, male; age: child, teen, youth, middle-age, senior; weight: under, healthy, over; clothing: workout, athletic, casual, dressy). 25,172 samples have all keypoints annotated, while 2,282 samples have at least one keypoint missed. For convenience, we call the dataset of 27,454 samples as the complete dataset, and that of 25,172 samples as the partial dataset. Follow the original paper, we divided the datasets into 10 train/test splits. In each split, samples of 4 videos were used for training and samples of the remaining 3 videos were used for testing.

We implemented a baseline approach which is similar to [1]. Our baseline uses a pre-trained very deep 16 CNN model [4] as the feature extractor. Similar to many recent works [9] [11], we take the 4,096 dimensional positive responses after FC6 and ReLU, and additionally perform the l2-norm normalization on them [13]. The fine-grained classification is then implemented by training one-vs-the-rest linear SVMs for each subcategory [14]. Our baseline is different from [1] in two aspects: (1) [1] uses the caffe reference model [15] of AlexNet [3] as the feature extractor, but our baseline uses the vgg very deep 16 model; (2) [1] extracts features from both the raw inputs and the warped inputs (warping is guided by the ground truth pose configurations), but our baseline extracts features only from the raw inputs.

We evaluated the following strategies with the baseline approach: super resolution (SR), patch dividing (PD), pose-wise classifier sharing with ground truth pose annotation (gPS) and graphical model (GM). We evaluated most of these strategies on both the complete dataset and the partial dataset. gPS was not evaluated on the complete dataset because some of the body part keypoints are not available. The evaluation of each approach was done on all the 10 train/test splits. The average accuracy (average accuracy is computed assuming that there is a uniform prior across the classes as in [1]) is reported in Table I, and the details are illustrated in Figure 3. Should note that the results that reported in [1] is also based on the complete dataset. However, because the detail information on the splits were not described, the splits used in our experiment may be different from the ones in [1].

B. Results and Discussions

From the results in Table I, we can confirm that all of the proposed strategies bring substantial improvements over the baseline approach. When integrate the strategies with the baseline approach individually, the overall classification performance on near all subcategories improves. Among the four strategies, two learning-free strategies show different properties. DR is quite efficient considering its simplicity, while SR is not quite stable from case to case. On the other hand, though learning is necessary, gPS and GM are quite stable and lead to moderate improvements.

We can also confirm that when integrate more than one proposed strategies to the baseline approach, the improvements become more obvious. Especially, without the PS strategy, the proposed approach outperforms the baseline with a margin of 3.32% on the complete dataset, and outperforms the baseline with a margin of 4.5% on the partial dataset. In case that the PS strategy is adopted, the margin becomes 4.73% on the partial dataset.

Compare to the approach in [1], the proposed approach outperforms it with a margin of 1.25% without using the PS strategy. As we can confirm from the results on the partial dataset, the PS strategy is able to bring additional improvements. The expected improvements over [1] should be more significant. Additionally, the proposed strategies can be integrated with [1] to directly improve it.

In Figure 3, we show the details of the classification results. We can see that for different subcategory classes, the proposed strategies show quite different effectiveness. However, when combine all of the strategies together, the improvements become more stable. We can also confirm that for the child class and the teen class in the age subcategory, combining all of the strategies leads to worse classification performance. We believe this is mainly due to the limitation of the CRP dataset. Since only 2.4% of the samples are labelled as child or teen, adopting multiple add-on strategies together on such an imbalanced dataset could easily lead to over-fitting. The issue is expected to be solved by preparing more annotated data of these two classes.

In the above experiments, the PS strategy was implemented using the ground truth pose annotations. In order to confirm how far we can go without ground truth labels, we implement and evaluated one more version of the PS strategy: PS with estimated pose configurations (ePS). We randomly selected one train/test split, and trained a deep CNN model for pose estimation using the approach described in [12]. The model is then used to estimate the pose configurations that are required by the ePS strategy on-the-fly. We report the results of two ePS variations, two gPS variations, and the baseline approach in Table II. We can confirm that though gPS outperforms ePS in both configurations, the differences are not obvious. Which suggest that the ePS strategy has sufficient robustness in practice.
In this paper, we proposed four add-on strategies for the fine-grained pedestrian classification task. These strategies have high flexibility and can be integrated with any end-to-end fine-grained pedestrian classification pipelines painlessly. We implemented and evaluated these strategies on the CRP dataset and confirmed that all of them lead to improvements over the baseline. We also confirmed substantial improvements over the state-of-the-art [1] when all strategies are combined together into a single pipeline.

In the future work, we would like to work on the following three aspects: (1) integrate these strategies with other fine-grained pedestrian classification pipelines, and do further evaluation on their effectiveness; (2) implement the strategies on problems of more subcategories and classes, and do further evaluation on their scalability; (3) study and improve the computational efficiency of these strategies from the perspective of real world applications.

ACKNOWLEDGMENT

This research is supported by the JSPS Grant-in-Aid for Scientific Research B (No. 26280057), the JSPS Grant-in-Aid for Challenging Exploratory Research (No.16K12460), and the JST Center of Innovation Program.

REFERENCES


TABLE I
THE EVALUATION RESULTS OF THE PROPOSED ADD-ON STRATEGIES.

<table>
<thead>
<tr>
<th>Method</th>
<th>Complete CRP Dataset</th>
<th>Partial CRP Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sex</td>
<td>age</td>
</tr>
<tr>
<td>Baseline</td>
<td>68.65</td>
<td>30.72</td>
</tr>
<tr>
<td>Baseline + SR</td>
<td>68.73</td>
<td>33.07</td>
</tr>
<tr>
<td>Baseline + DR</td>
<td>72.30</td>
<td>31.68</td>
</tr>
<tr>
<td>Baseline + gPS</td>
<td>70.30</td>
<td>31.24</td>
</tr>
<tr>
<td>Baseline + GM</td>
<td>68.65</td>
<td>32.36</td>
</tr>
<tr>
<td>D. Hall, et al. [1]</td>
<td>78</td>
<td>35</td>
</tr>
<tr>
<td>Proposed w/o gPS</td>
<td>71.82</td>
<td>32.68</td>
</tr>
<tr>
<td>Proposed with gPS</td>
<td>75.89</td>
<td>31.72</td>
</tr>
</tbody>
</table>

TABLE II
COMPARISON BETWEEN ePS AND gPS. (%)
Fig. 3. Details of the classification results. The number under the class name denotes the total number of images of that class.