Robust Pedestrian Attribute Recognition for an Unbalanced Dataset using Mini-batch Training with Rarity Rate

Hiroshi Fukui[1], Takayoshi Yamashita[1], Yuji Yamauchi[1], Hironobu Fujiyoshi[1] and Hiroshi Murase[2]

***Abstract*— Pedestrian attributes are significant information for Advanced Driver Assistance System(ADAS). Pedestrian attributes such as body poses, face orientations and open umbrella are meant action or state of pedestrian. In general, this information is recognized using independent classifiers for each task. Performing all of these separate tasks is too time-consuming at the testing stage. In addition, the processing time increases with the number of tasks. To address this problem, multi-task learning or heterogeneous learning is able to train a single classifier to perform multiple tasks. In particular, heterogeneous learning is able to simultaneously train regression and recognition tasks, because reducing both training and testing time. However, heterogeneous learning tends to result in a lower accuracy rate for classes with a few training samples. In this paper, we propose a method to improve the performance of heterogeneous learning for such classes. We introduce a rarity rate based on the importance and class probability of each task. The appropriate rarity rate is assigned to each training sample. Thus, the samples in a mini-batch for training a deep convolutional neural network are augmented by this rarity rate to focus on the class with a few samples. Our heterogeneous learning approach with the rarity rate attains better performance on pedestrian attribute recognition, especially for classes representing open umbrellas.**

I. INTRODUCTION

In an Advanced Driver Assistance System (ADAS) [1], object recognition on vehicle camera assists the driver in decision-making under hazardous conditions. Generally, the ADAS includes pedestrian or vehicle detection and traffic sign recognition. To prevent collisions between vehicles and pedestrians, these detection technologies are a key function within the ADAS.

The combination of the histogram of oriented gradient (HOG) and the support vector machine (SVM) is a common approach for pedestrian detection [2]. The HOG feature focuses on the gradient of a local region and is robust to small variations in pose. Following Dalal work, several related pedestrian detection methods have been proposed [3] [4] [5] [6] [7]. With the proliferation of deep learning, the convolutional neural network (CNN) has become a common classifier for pedestrian detection [8].

To advance the current state of ADAS, pedestrian attribute recognition is part of the key functions for improvement. Pedestrian attributes are significant in supporting intelligent ADAS decisions. For example, the orientations of a pedestrian's body and face are noticeable attributes. Such attributes are used to predict pedestrian behavior (e.g., aiding the ADAS in predicting that the pedestrian will suddenly run in front of the vehicle). As another example, collisions between vehicles and pedestrians occur more frequently on rainy days than sunny days [9]. Thus, the attribute signifying weather a pedestrian has an open umbrella is important in predicting and preventing these traffic conflict.

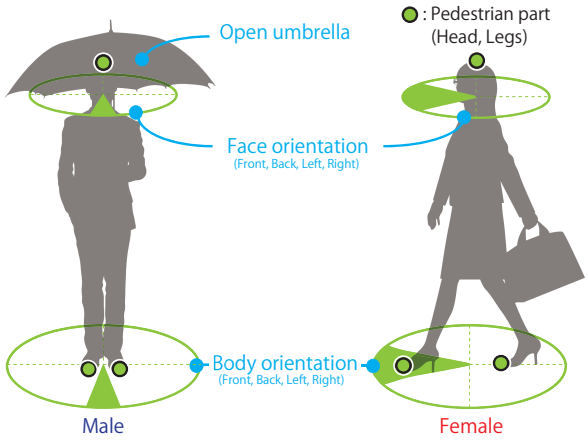

Fig. 1. Recognizing pedestrian attributes using heterogeneous learning

The common approach is to train a classifier for each task such as recognizing the body or face orientation and other attributes. Unfortunately, this is inefficient because the computational cost of training and testing increases with the number of tasks. To address this problem, multi-task learning [10] trains a single classifier to carry out multiple tasks. A CNN trained using multi-task learning has units outputting the recognition results corresponding to each task. Thus, a single neural network classifies multiple tasks simultaneously and the computational cost does not vary according to the number of tasks. If it recognizes multiple heterogeneous tasks which regression tasks and recognition tasks, heterogeneous learning is recognized their tasks. Heterogeneous Learning can train the multiple heterogeneous tasks by selecting error function of regression and recognition. Several methods have proposed the use of heterogeneous learning [12] [13]. In this paper, we consider multiple pedestrian attributes including simultaneous recognition and regression tasks using heterogeneous learning, as shown in Fig. 1.

To train the classifier with heterogeneous learning, it is necessary to prepare the dataset whereby samples have multiple labels for each task. Typically, the number of samples for each class of each task is unbalanced and it is difficult to construct a dataset such that each class of each task has an equal number of samples. The performance of a CNN trained by heterogeneous learning tends to significantly

[1]Chubu University, 1200 Matsumoto cho, Kasugai, Aichi, Japan fhiro@vision.cs.chubu.ac.jp
[2]Nagoya University, Furo cho, Chikusa ku, Nagoya, Aichi, Japan

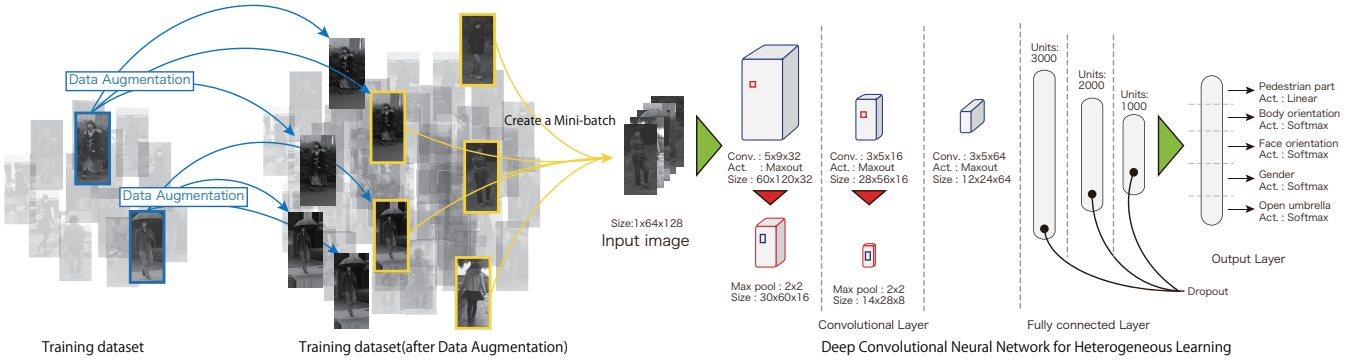

Fig. 2. Deep convolutional neural network for heterogeneous learning

deteriorate for classes with few samples. This problem is inherent in the training process of the CNN. Mini-batch is a common approach used to update the CNN parameters by backpropagation with the error of the subset which contains a few training samples. In traditional mini-batch approaches, the training samples are chosen randomly from the training dataset. The probability of choosing a training sample from a class with few samples is lower than the probability of choosing a sample from a class with many samples. Thus, the performance of classes which consist a few samples is likely to suffer from difficult to choices the training sample equally.

In this paper, we propose a new mini-batch selection approach that introduces the rarity rate of training samples to alleviate the above problem. First, we use the rarity rate to prepare the subset to balance between the common and rare sample accordingly. The rarity rate is based on the ratio of the number of samples in the classes corresponding to each task. The probability of choosing the samples from a class with few samples is increased in the subset as compared with traditional mini-batch. The samples chose according to the rarity rate are subjected to data augmentation to increase variation. Then, we select mini-batch candidates randomly from the augmented subset and select one mini-batch that appropriately represents the rare samples. The heterogeneous learning CNN using our proposed mini-batch creation method improves the recognition performance for classes with a few samples.

II. Pedestrian Attribute Recognition Using a Heterogeneous Learning CNN

We categorize related work into pedestrian attribute recognition and heterogeneous learning. In the following subsections, we describe the related methods in these categories and then further discuss the problems with existing heterogeneous learning CNN methods as applied to pedestrian attribute recognition.

A. Related pedestrian attribute recognition work

Pedestrian detection is part of an important ADAS function. Dalal proposed an impressive pedestrian detection method that employs HOG and SVM [2]. The HOG extracts the gradient value rather than pixel values and is thus robust to pedestrian appearance variations. The features derived using HOG has been proposed in many publications [3] [4] [5] [6].

Since Krizhevsky successfully achieved object recognition using CNN [14], the deep learning architecture has become widely applied in computer vision and pedestrian detection in particular [8]. Hosang used AlexNet, a common deep learning architecture, for pedestrian detection and achieved impressive performance [8]. The aforementioned work also analyzed the performance for a number of training samples.

Pedestrian attribute recognition is also one of the most key function in ADAS and is a significant part of reducing collisions between vehicles and pedestrians. Ricci proposed a method to estimate the body and face orientation from RGB and stereo images [15]. This method employs a dynamic Bayesian network for estimation and each input region is extracted using tracking on a sequence of stereo images. Bearman proposed human pose estimation and joint position detection using CNN [16]. While this method attains high pose estimation performance for complex human poses, the networks for body pose detection and joint position estimation are constructed individually.

B. Heterogeneous learning

Performing recognition or estimation for multiple tasks requires the construction of classifiers corresponding to each task, as shown in Fig. 2. This is time-consuming during training and testing, and the computation time increases with the number of tasks. One of the methods developed to address this problem is heterogeneous learning. Heterogeneous learning performs multiple tasks in a single network. A CNN trained for heterogeneous learning has units that output the recognition results corresponding to each task. The computational cost does not directly depend on the number of tasks. Zhang proposed a method to perform multiple tasks such as facial point estimation, gender classification, face orientation estimation, and glasses detection [13]. While the method estimates multiple task, its main purpose is to improve the performance of the primary task, such as facial point detection. It thus assigns weighted loss functions to each task. When the loss decreases sufficiently, the training of the task is terminated earlier to avoid over-fitting to a specific task.

During the pre-processing to train the CNN, the training samples are augmented. Data augmentation is a common strategy to increase the number of training samples using translation, scaling, and rotation. After data augmentation, M training samples are chosen randomly to form the mini-batch. In mini-batch training, the error E is calculated and backpropagated to update the parameters of the network. At each backpropagation [18] iteration, the samples in the mini-batch are selected randomly from the augmented dataset. When the CNN is trained with heterogeneous learning, the recognition and regression tasks are combined in a single network and each task has an independent loss function. The cross entropy in Eq.(1) and the mean squared-error in Eq.(2) are employed as the loss functions of the recognition and regression tasks, respectively.

$$E_{m,t}^{Classification} = -\mathbf{y} \log \mathbf{o} \tag{1}$$

$$E_{m,t}^{Regression} = \| \mathbf{y} - \mathbf{o} \|_2^2 \tag{2}$$

Note that, $\mathbf{o}$ and $\mathbf{y}$ mean labels and response for each task, respectively. The errors $E_{m,t}$ of the sample m for all tasks $\{t|1,\ldots,T\}$ are accumulated and propagated once per iteration in Eq.(3). The parameters $\boldsymbol{W}$ of the CNN are updated by using the differential of the accumulated error with the training coefficient η. In this paper, we use a heterogeneous learning CNN for pedestrian part detection, body orientation, face orientation, gender, and the presence of an open umbrella, as illustrated in Fig 1.

$$\begin{aligned} \mathbf{W} &\leftarrow \mathbf{W} + \Delta\mathbf{W} \\ &= \mathbf{W} - \eta \frac{\partial \sum_{m=1}^{M} \sum_{t=1}^{T} E_{m,t}}{\partial \mathbf{W}} \end{aligned} \tag{3}$$

C. Drawbacks of the heterogeneous learning CNN

A CNN trained by heterogeneous learning typically shows poor performance for classes with few samples. Figure 3 shows the relationship between the ratio of the number of training samples and the accuracy for each task. The accuracy for the classes with few training samples is lower than for those with more training samples.

This happens characteristic of the training process of the CNN. The samples belonging to classes with many samples are selected in the mini-batch and the error is propagated frequently. However, the error of a class with a few samples is rarely propagated. As a result, the performance of these small classes deteriorates. This problem corresponds to the different sample distributions in each class of each task.

III. Proposed Method

To improve the performance for classes with few samples, we propose a method using a rarity rate assigned to each sample. The samples forming the mini-batch are choosing based on rarity rate. Note that we refer to samples from small classes as rare samples and samples from large classes as common samples. In conventional mini-batch creation, the performance for the rare classes is worse because of the random choice of training samples. The proposed method improves the performance for these classes by increasing the number of choice times for rare samples by using the rarity rate.

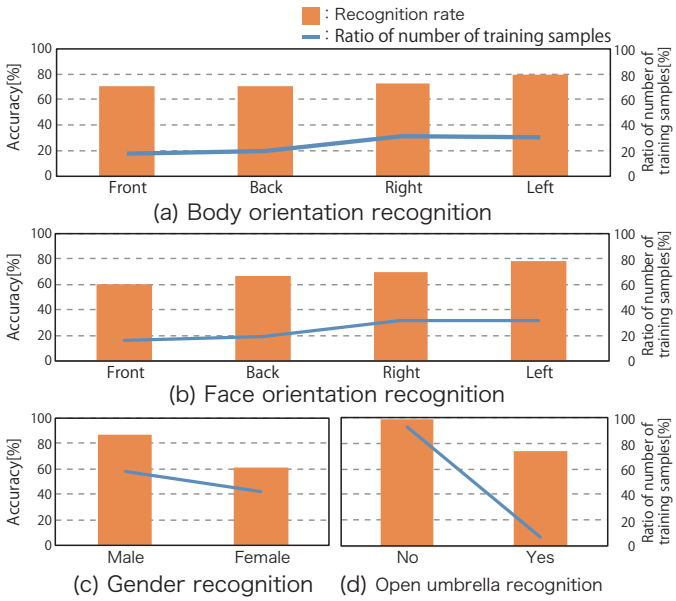

Fig. 3. Relationship between the number of training samples and accuracy for each task

First, we define the rarity rate for each training sample $\{n|1,\ldots,N\}$. Then, the training samples n are augmented by the rarity rate. Using the rarity rate, the rare samples are augmented to many samples. In contrast, the augmentation of common samples is suppressed to a few samples.

After data augmentation, the samples are chosen randomly to form the mini-batch. We create several mini-batches as mini-batch candidates and select one that has an appropriate sample balance. The network is trained and its parameters are updated using the selected mini-batch. New mini-batch candidates are created and selected at each iteration. The following subsections provide detailed information about these algorithms.

A. Definition of rarity rate

The rarity rate of a training sample n is a quantitative value signifying the rarity of the corresponding class in each task. Figure 4 shows the process to assign the rarity rate to a training sample. Each training sample has labels for each task. In this case, choosing training sample has four labels which are left body orientation, left face orientation, female gender, and no umbrella. These ratios of the training samples in each class for the task t are calculated from these attribute labels of all training samples. The rarity rate R_n of a training sample is defined by Eq.(4).

$$R_n = \sum_{t=1}^{T-1} (1 - p_{n,t}) \cdot \sqrt{1 - \frac{p_t^{min}}{p_t^{max}}} / T \tag{4}$$

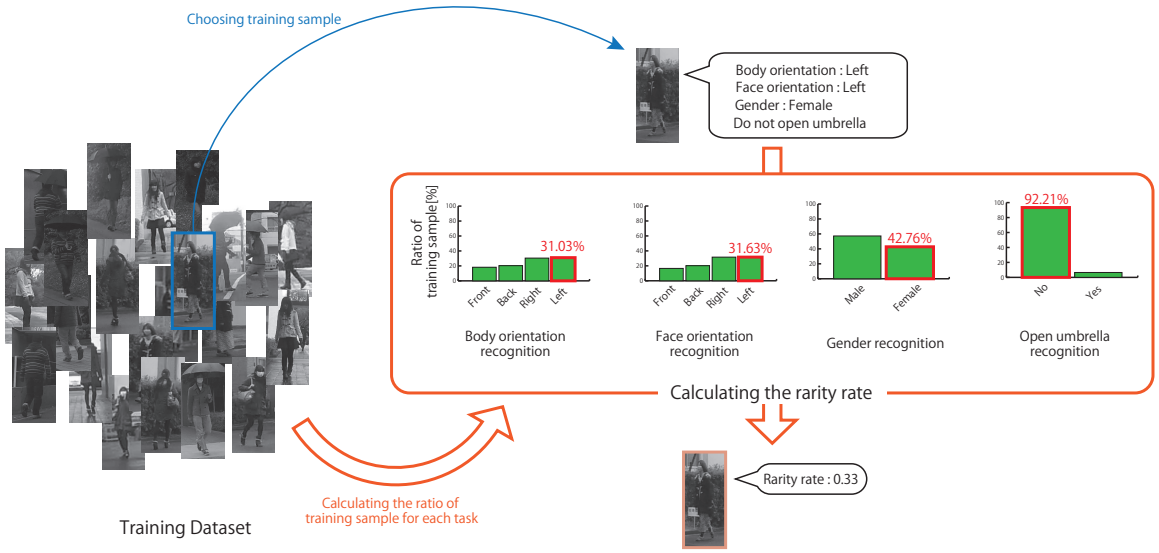

Fig. 4. Definition of the rarity rate

Note that $p_{n,t}$ is the ratio of training samples in the corresponding class for task t. p_t^{min} and p_t^{max} are the minimum and maximum ratios of training samples, respectively.

The first term in Eq.(4) indicates the rarity rate of the corresponding class. The second term indicates the deviation between classes in the task t.

If a training sample belongs to a class with a few training samples, the rarity rate of the first term increases. In contrast, if the training sample belongs to a class with many training samples, the rarity rate of the first term decreases. The deviation of the training samples of the task is obtained from the ratios of training samples in each class. It is defined by the classes that have the minimum and maximum amounts of training samples. To suppress the deviation between classes, the ratio of the class with the most training samples is normalized to one.

B. Data augmentation considering the rarity rate

Because the random choice probability of rare samples is lower than that of common samples, traditional mini-batch creation severely under-represents rare samples. While data augmentation can increase the number of samples, both rare and common samples are increased equally using a conventional approach. Thus, our proposed method applies data augmentation based on the rarity rate. We thereby augment rare samples and suppress the augmentation of common samples.In addition, we do not do transformations such as mirroring which changed especially belonging class from data augmentation. This can result in an augmented dataset that includes equal numbers of training samples for each class for each task. The augmented number of samples S_n is defined using the rarity rate R_n as shown in Eq.(5).

$$S_n = R_n \cdot A + 1 \tag{5}$$

A is the maximum augmentation number for increasing the training samples by data augmentation. It suppresses the augmentation of common samples by the number of augmented samples corresponding to the rarity rate. The probability of rare sample choice is thus increased.

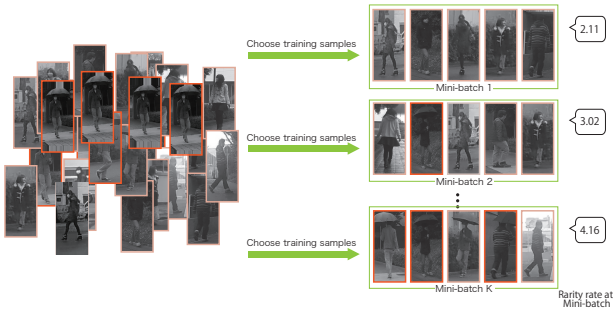

Fig. 5. Creating the candidate mini-batches

C. Selecting the mini-batch candidate

To select the mini-batch with the appropriate sample balance, we create several mini-batch candidates. Each mini-batch candidate is constructed by choosing samples from the augmented dataset described in the previous subsection. The rarity rate for mini-batch candidate R_k^B is defined from the total rarity rates of the chose sample R_m as shown in Eq.(6). K mini-batch candidates are considered preparing with the rarity rate.

$$R_k^B = \sum_{m=1}^{M} R_m \tag{6}$$

Mini-batch candidates are sorted by their rarity rates R_k^B and the one with the median rate is selected. Mini-batch candidates with high rarity rates include many rare samples, and mini-batch candidates with low rarity rates include many common samples. The mini-batch with the median rarity rate balances between rare and common samples.

IV. Experiments

A. Overview

We evaluate our method for multiple tasks with datasets that are unbalanced in the number of samples for each class.

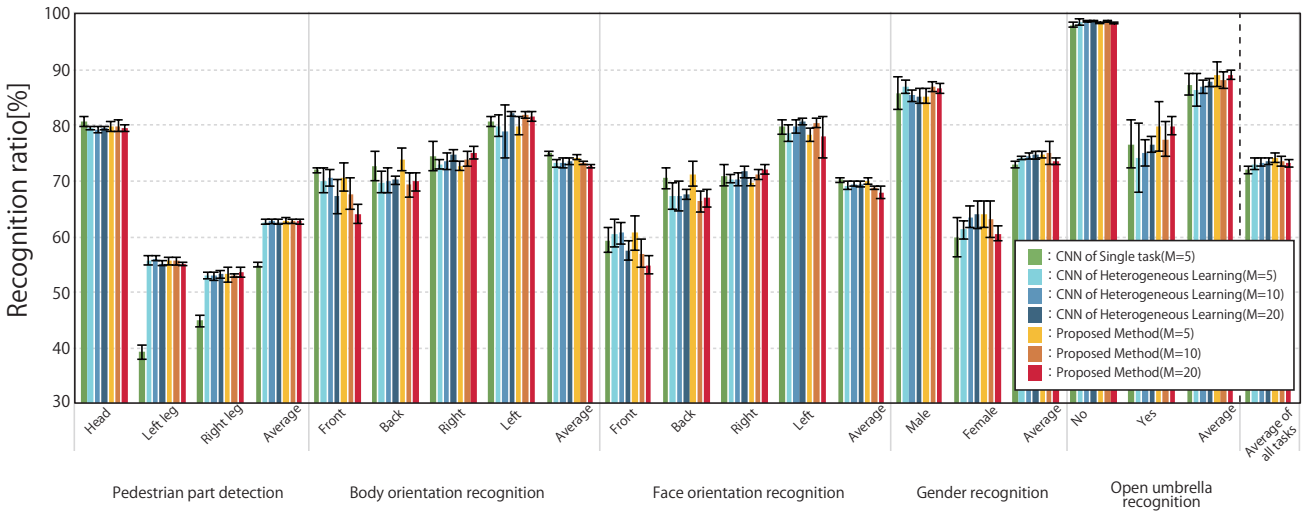

Fig. 6. Comparison between conventional methods and proposed method

In the experiments, we compare the performance of proposed method and conventional methods. Note that, we preliminarily evaluate the parameter of our method are the maximum number of augmentation samples A and the number of mini-batch candidate K. Once the optimal parameters are used, we compare the performance of our method with conventional methods. The comparison methods are:

- Recognition of each task using individual CNNs,
- Recognition of all tasks by a single CNN trained by heterogeneous learning,
- Recognition of all tasks by the proposed method.

To compare performance, we evaluate all methods with varying mini-batch sizes M of values $\{5,\ 10,\ 20\}$. The comparison dataset consists of 82,364 pedestrian images that were taken from a vehicle camera. In addition, we divide training samples of 45,581 and test samples of 36,783 in this dataset. Each pedestrian image is labeled for five tasks: pedestrian part position, body orientation, face orientation, gender, and open umbrella detection, as illustrated in Fig.1. These ratios of training samples in each task are shown in Fig.5. The class with the least samples is "open umbrella" in all of the tasks. The performance of each task except pedestrian separate detection is evaluated according to the recognition rate. For pedestrian part detection, we evaluate the localization error as a fraction of the head-to-legs distance (this is invariant with respect to the actual size of the images). A point has been correctly detected if the pixel error is lower than 10% of the head-to-legs distance.

In this experiment, we employ a CNN that consists of three convolutional layers and three fully connected layers, as shown in Fig. 2. The total number of iterations to update the parameters is 500,000, and the training ratio η is set to 0.001. We evaluate each method five times to smooth variability and calculate the standard deviation. In addition, we set the same initial parameters for all methods.

B. Experimental result

In Fig. 6, we compare the performance of our proposed method with single task learning and heterogeneous learning. We show the recognition rates as the parameter M varies within $\{5,\ 10,\ 20\}$. In our method, we set to $A = 25$ and $K = 15$ why are best performance in preliminarily experiments which changed the parameters A and K, respectively.

Figure 6 shows that the CNN with heterogeneous learning is superior the single task approach for pedestrian part detection. In particular, the detection performance of the CNN with heterogeneous learning improves by more than 10% for both legs as compared with the single task approach. The body orientation attribute is supported by part detection, especially the legs, because the part position and body orientation recognition can be performed simultaneously. The proposed method improves the performance for rare classes such as "open umbrella", "body orientation is front", "face orientation is front", and "female gender" by applying the rarity rate to heterogeneous learning. When we evaluate the various values of the mini-batch size M, the conventional heterogeneous learning achieves its best performance with M equal to 20. Our method achieves its best performance with M equal to 5.

In Fig. 7, we show the recognition results. The circles at the head and foots denote the orientation of the face and body, respectively. The green points are detected part positions in the head and legs. The color of the bounding box indicates gender (blue is male and red is female). In addition, when an open umbrella is detected, an umbrella icon is shown on the right-bottom. As shown in Fig.7, both conventional heterogeneous learning and the proposed method can recognize multiple tasks for various pedestrian poses. However, the conventional method has difficulty recognizing classes with few samples such as "female pedestrian", "body orientation is front", "face orientation is front", and "open umbrella". The proposed method improves the performance for these classes. In particular, the class "open umbrella" is significantly improved with respect to the conventional method.

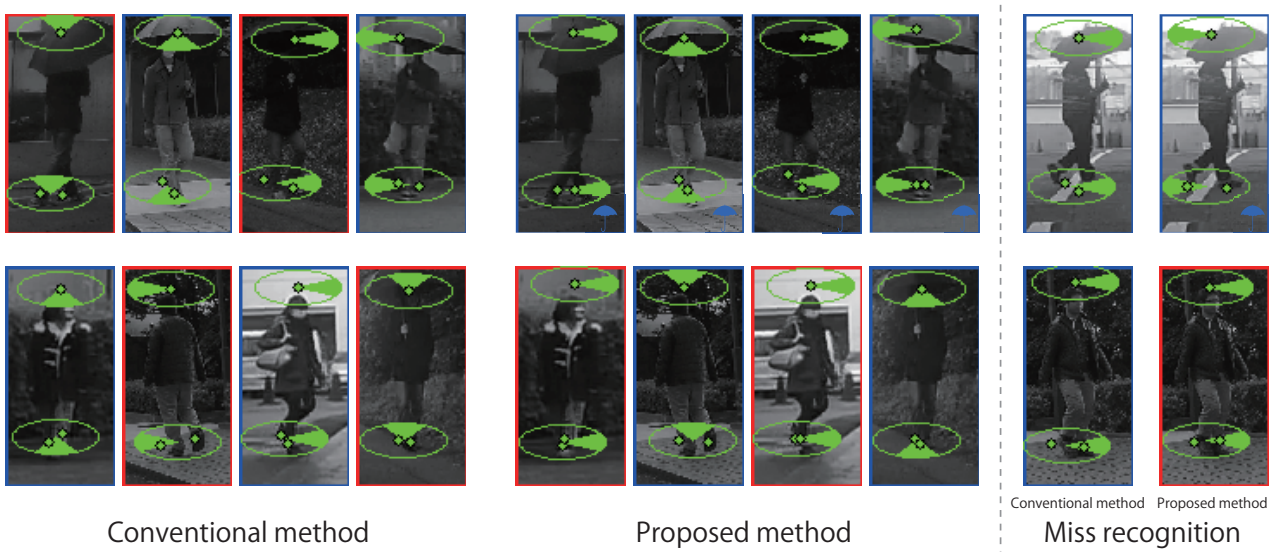

Fig. 7. Comparison of example pedestrian attribute recognition using the conventional method and proposed method

V. Conclusion

In this paper, we propose a method to improve the performance of heterogeneous learning of multi-task pedestrian attribute recognition for classes with few samples. We assign a rarity rate to each training sample to determine the number of augmentations such that rare samples are augmented more. In addition, we create multiple mini-batch candidates from the augmented datasets and the candidate that balances appropriately balances between common and rare samples is selected as the training mini-batch. As a result, the proposed method improves the recognition performance for classes with few sample. By introducing the rarity rate during data augmentation, the proposed method reduces the number of training samples by half.

ACKNOWLEDGMENT

This research is partially supported by Center of Innovation Program from Japan Science and Technology Agency, JST.

References

[1] D. Geronimo, A. M. Lopez, and A. D. Sappa, "Survey of Pedestrian Detection for Advanced Driver Assistance Systems", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32, No. 7, pp.1239-1258, 2010.

[2] N. Dalal, and B. Triggs, "Histograms of oriented gradients for human detection", Computer Vision and Pattern Recognition, 2005.

[3] P. Felzenszwalb, D. McAllester, and D. Ramaman, "A Discriminatively Trained, Multi scale, Deformable Part Model", Computer Vision and Pattern Recognition 2008.

[4] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP Human Detection with Partial Occlusion", International Conference on Computer Vision, 2009.

[5] W. Nam, B. Han, and J. H. Han, "Improving Object Localization Using Macrofeature Layout Selection", International Conference on Computer Vision Workshop on Visual Surveillance, 2011.

[6] J. Marin, D. Vazquez, A. M. López, J. Amores, and B. Leibe, "Random Forests of Local Experts for Pedestrian Detection", International Conference on Computer Vision, 2012.

[7] H. Fukui, T. Yamashita, Y. Yamauchi, H. Fujiyoshi, and H. Murase, "Pedestrian Detection Based on Deep Convolutional Neural Network with Ensemble Inference Networks", IEEE Intelligent Vehicle Symposium, 2015.s

[8] J. Hosang, M. Omran, R. Benenson, and B. Schiele, "Taking a Deeper Look at Pedestrian", Computer Vision and Pattern Recognition, 2015.

[9] H. Brodsky, and A. S. Hakkert, "Risk of a road accident in rainy weather", Accident Analysis & Prevention, Vol. 20, Iss. 3, pp. 161-176, 1988.

[10] A. Andreas, E. Theodoros, and P. Massimiliano, "Convex Multi-task Feature Learning", Kluwer Academic Publishers, Vol. 73, No. 3, pp. 243-272, 2008.

[11] X. Yang, S. Kim, and F. P. Xing, "Heterogeneous Multi-task Learning with Sparsity Constrain", Advances in Neural Information Processing Systems 22, 2009.

[12] S. Li, Z. Q. Liu, and A. B. Chan, "Heterogeneous Multi-task Learning for Human Pose Estimation with Deep Convolutional Neural Network", Computer Vision and Pattern Recognition, 2014.

[13] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial Landmark Detection by Deep Multi-task Learning", in Proceedings of European Conference on Computer Vision, 2014.

[14] A. Krizhevsky, S. Ilva, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Network", Advances in Neural Information Processing System 25, pp.1097-1105, 2012.

[15] E. Ricci, J. Varadarajan, R. Subramanian, S. Rota-Bulo, N. Ahuja, and O, Lanz, "Uncovering Interactions and Interactors: Joint Estimation of Head, Body Orientation and F-formations from Surveillance Videos", Association for Computing Machine, 2015.

[16] A. Bearman, and C. Dong, "Human Pose Estimation and Activity Classification Using Convolutional Neural Networks", CS231n Course Project Reports, 2015.

[17] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian Detection: An Evaluation of the State of the Art", Pattern Analysis and Machine Intelligence, Vol. 34, 2012.

[18] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors", Neurocomputing, pp. 696-699, 1988.