PAPER *Special Section on Machine Vision and its Applications*

Training of CNN with Heterogeneous Learning for Multiple Pedestrian Attributes Recognition Using Rarity Rate

Hiroshi FUKUI[†a)], **Takayoshi YAMASHITA**[†b)], **Yuji YAMAUCHI**[†c)], **Hironobu FUJIYOSHI**[†d)], ***and* Hiroshi MURASE**[††e)], ***Members***

SUMMARY Pedestrian attribute information is important function for an advanced driver assistance system (ADAS). Pedestrian attributes such as body pose, face orientation and open umbrella indicate the intended action or state of the pedestrian. Generally, 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 increasing number of tasks. To address this problem, multi-task learning or heterogeneous learning is performed to train a single classifier to perform multiple tasks. In particular, heterogeneous learning is able to simultaneously train a classifier to perform regression and recognition tasks, which reduces both training and testing time. However, heterogeneous learning tends to result in a lower accuracy rate for classes with 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 according to this rarity rate to focus on the classes with a few samples. Our heterogeneous learning approach with the rarity rate performs pedestrian attribute recognition better, especially for classes representing few training samples.

key words: *pedestrian attributes recognition, heterogeneous learning, rarity rate*

1. Introduction

In an advanced driver assistance system (ADAS) [1], object recognition using the vehicle camera assists the driver in decision-making under hazardous conditions. Generally, the ADAS functions include pedestrian or vehicle detection and traffic sign recognition. To prevent collisions between vehicles and pedestrians, these detection technologies are a key function for an 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 the work of Dalal, several related pedestrian detection methods have been proposed [3]–[6]. With the proliferation of deep learning, the convolutional neural network (CNN) has become a common classifier for pedestrian detection [7]–[9].

To improve the ADAS, pedestrian attribute recognition is one of the key functions. Pedestrian attributes are important for supporting intelligent ADAS decisions. For example, the orientation 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, more collisions between vehicles and pedestrians occur on rainy days than on sunny days [10]. Thus, the attribute signifying whether a pedestrian has an open umbrella is important in predicting and preventing these traffic collisions.

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 increasing number of tasks. To address this problem, multi-task learning [11], which trains a single classifier to carry out multiple tasks, is used. 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 consist of regression tasks and recognition tasks, heterogeneous learning can recognize their tasks. Heterogeneous learning can train a classifier to perform multiple heterogeneous tasks by selecting the error function of regression and recognition.

Several methods have proposed the use of heterogeneous learning [13], [14]. In this paper, we consider detecting multiple pedestrian attributes using simultaneous recognition and regression tasks with heterogeneous learning, as shown in Fig. 1.

To train the classifier with heterogeneous learning, it is necessary to prepare a dataset where 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 deteriorate for classes with few samples. This problem is inherent in the training process of the CNN. Mini-batch is a

Manuscript received June 13, 2017.
Manuscript revised November 30, 2017.
Manuscript publicized February 16, 2018.

[†] The authors are with the Chubu University, Kasugai-shi, 487–0027 Japan.

[††] The author is with the Nagoya University, Nagoya-shi, 464–0814 Japan.

a) E-mail: fhiro@mprg.cs.chubu.ac.jp
b) E-mail: yamashita@chubu.ac.jp
c) E-mail: yuu@isc.chubu.ac.jp
d) E-mail: hf@chubu.ac.jp
e) E-mail: murase@is.nagoya-u.ac.jp

DOI: 10.1587/transinf.2017MVP0001

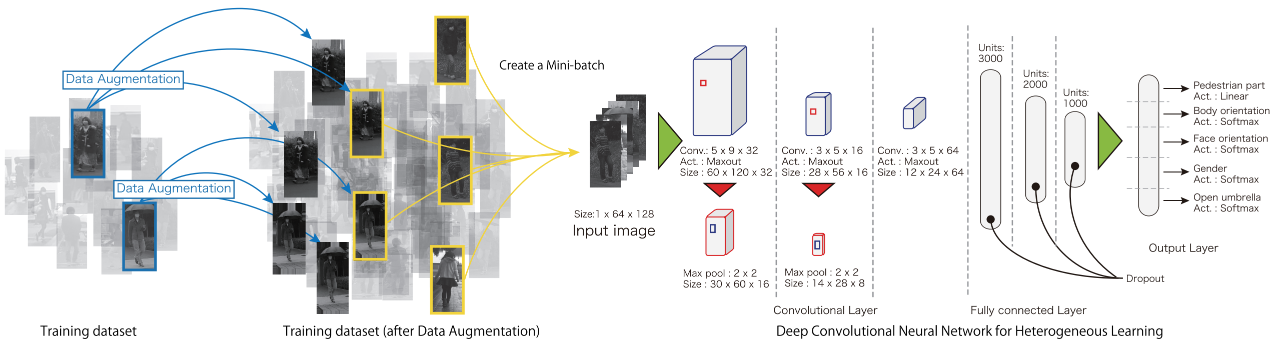


Fig. 2 Deep convolutional neural network for heterogeneous learning

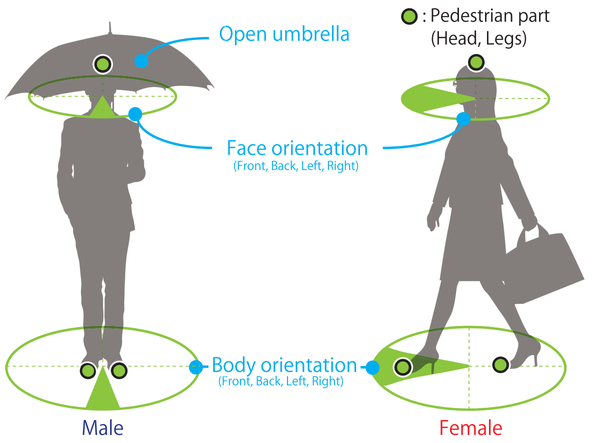


Fig. 1 Recognizing pedestrian attributes using heterogeneous learning

common approach used to update the CNN parameters by backpropagation with the error of the subset that contains 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 training sample from a class with many samples. Thus, the performance of classes that consist of few samples is likely to be worse because it is difficult to choose the training samples 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 reach a balance between the common and rare samples 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 compared with using a traditional mini-batch. The samples chosen 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 with CNN using our proposed mini-batch creation method improves the recognition performance for classes with few samples.

2. Pedestrian Attribute Recognition Using Heterogeneous Learning CNN

We categorize the 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.

2.1 Related Pedestrian Attribute Recognition Work

Pedestrian detection is one of the important ADAS functions. Dalal proposed an impressive pedestrian detection method that uses the HOG and SVM [2]. The HOG extracts the gradient value rather than pixel values, thus, it is robust to pedestrian appearance variations. The features derived using the HOG have been proposed in many publications [3]–[6].

Since Krizhevsky successfully achieved object recognition using a CNN [20], the deep learning architecture has been widely applied in computer vision and pedestrian detection in particular [7]–[9]. Hosang used AlexNet, a common deep learning architecture, for pedestrian detection and achieved impressive performance [9]. The aforementioned work also analyzed the performance for a number of training samples.

Pedestrian attribute recognition is also one of the most key functions in an ADAS and is a significant factor in reducing collisions between vehicles and pedestrians. Ricci proposed a method to estimate the body and face orientation from RGB and stereo images [22]. This method uses 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 a CNN with heterogeneous learning [23]. 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.

2.2 Heterogeneous Learning

Performing recognition or estimation for multiple tasks requires the construction of classifiers corresponding to each task. This is time-consuming during training and testing, and the computation time increases with increasing number of tasks. One of the methods developed to address this problem is heterogeneous learning. One of the methods developed to address this problem is multi-task learning. Multi-task learning performs multiple tasks in a single network. Heterogeneous learning, a specific type of multi-task learning, can train multiple regression tasks and recognition 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 [14]. While this 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 overfitting to a specific task. This establishes the effectiveness of heterogeneous learning for attribute recognition.

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 and scaling. 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 [25] 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 used as the loss functions of the recognition and regression tasks, respectively.

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

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

Note that, **o** and **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 $\boldsymbol{\theta}$ 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 to classify pedestrian attributes such as pedestrian part detection, body orientation, face orientation, gender, and the presence of an open umbrella, as illustrated in Fig 1.

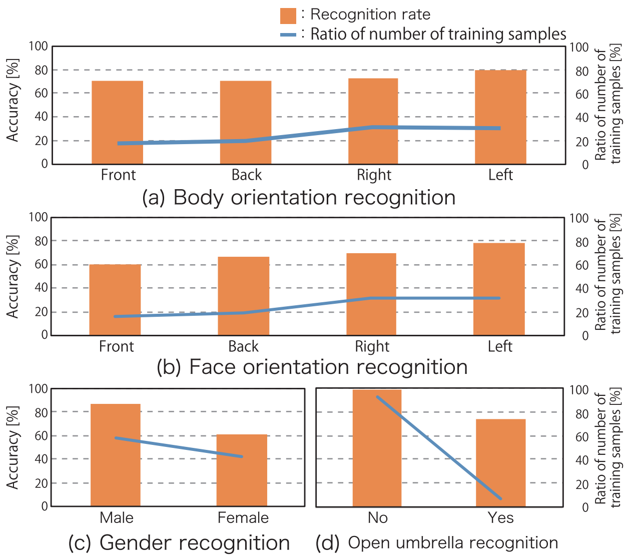


Fig. 3 Relationship between the number of training samples and accuracy for each task

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

2.3 Drawbacks of the CNN with Heterogeneous Learning

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

This is a 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 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.

Studies on imbalanced data in deep learning have been conducted extensively [15]–[19]. In previous works [17]–[19], a weighted loss function was applied to a CNN in training to address the problem of unbalance data. Weight was defined in these studies using the number of samples of particular class. Huang proposed quintuplet learning with the associated triple-header loss that preserves locality across clusters and discrimination between classes [16]. These previous works are defined the weight of the cost function or of the minority class and majority class using the number of training samples in the training dataset. However, it is difficult to adopt this number of training samples into training because CNN are trained using a few sample mini-batches in one back-propagation. We represented the rarity rate with the frequency of training samples, unlike in the work by Huang, in which it is work of represented by minority and

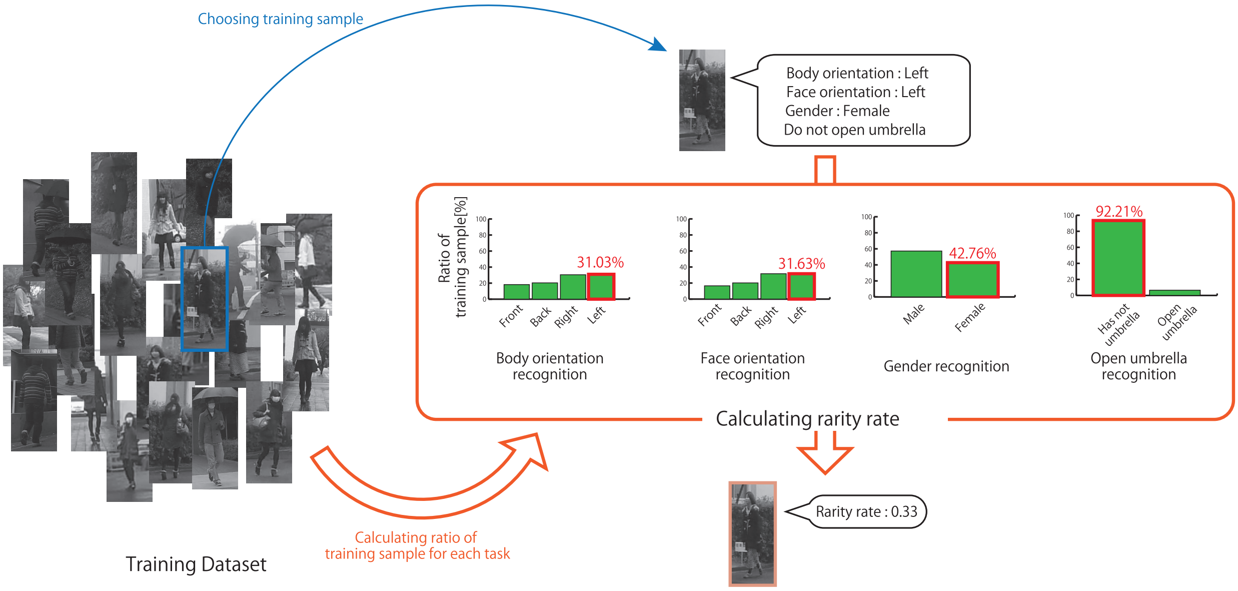


Fig. 4 Definition of the rarity rate

majority class. We were able to efficiently select a few sample of mini-batches using this rarity rate.

3. 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 chosen 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 that for rare samples are chosen by using the rarity rate.

First, we define the rarity rate for each training sample $\{n|1,\ldots,N\}$. Then, the training samples n are augmented according to the rarity rate. By using the rarity rate, the rare samples are augmented to give many samples. In contrast, the augmentation of common samples is suppressed so that there are only a few common 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.

3.1 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, the chosen 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 for all training samples. The rarity rate R_n of a training sample is defined by Eq. (4).

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

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 the classes in the task t.

If a training sample belongs to a class with 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 number of training samples. To suppress the deviation between classes, the ratio of the class with the most training samples is normalized to one.

3.2 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, the number of both rare and common samples are increased equally

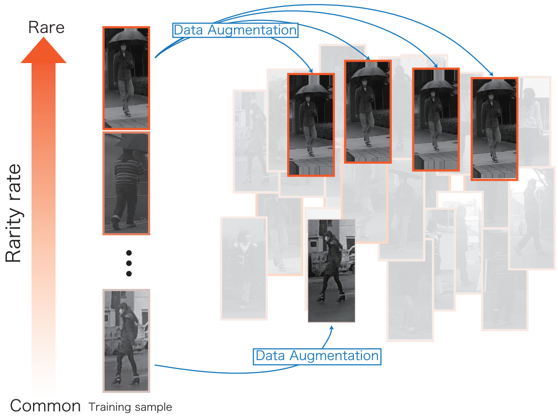


Fig. 5 Data augmentation considering rarity rate

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, as shown in Fig. 5. In addition, we did not perform transformations such as mirroring, which change in particular the class that the sample belonged to due to data augmentation. This could 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 number of training samples by data augmentation. The augmentation of the common samples by the number of augmented samples corresponding to the rarity rate is suppressed. The probability of choosing a rare sample choice is thus increased.

3.3 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 at random from the augmented dataset described in the previous subsection. The rarity rate for mini-batch candidate R^B_k is defined based on the total rarity rates of the chose sample R_m as shown in Eq. (6). K mini-batch candidates are considered to have been prepared with the rarity rate.

$$R^B_k = \sum_{m=1}^{M} R_m \tag{6}$$

Mini-batch candidates are sorted by their rarity rates R^B_k 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 has a balance of rare and common samples.

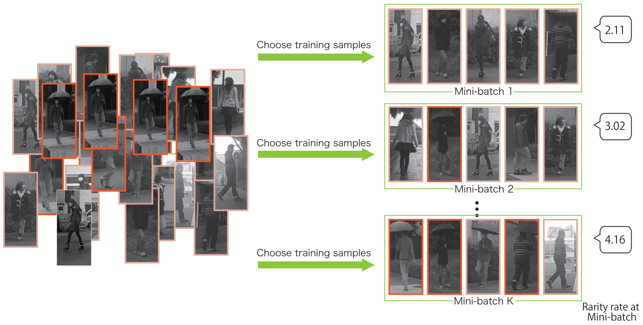


Fig. 6 Creating candidate mini-batches

Algorithm 1 Training algorithm of proposed method

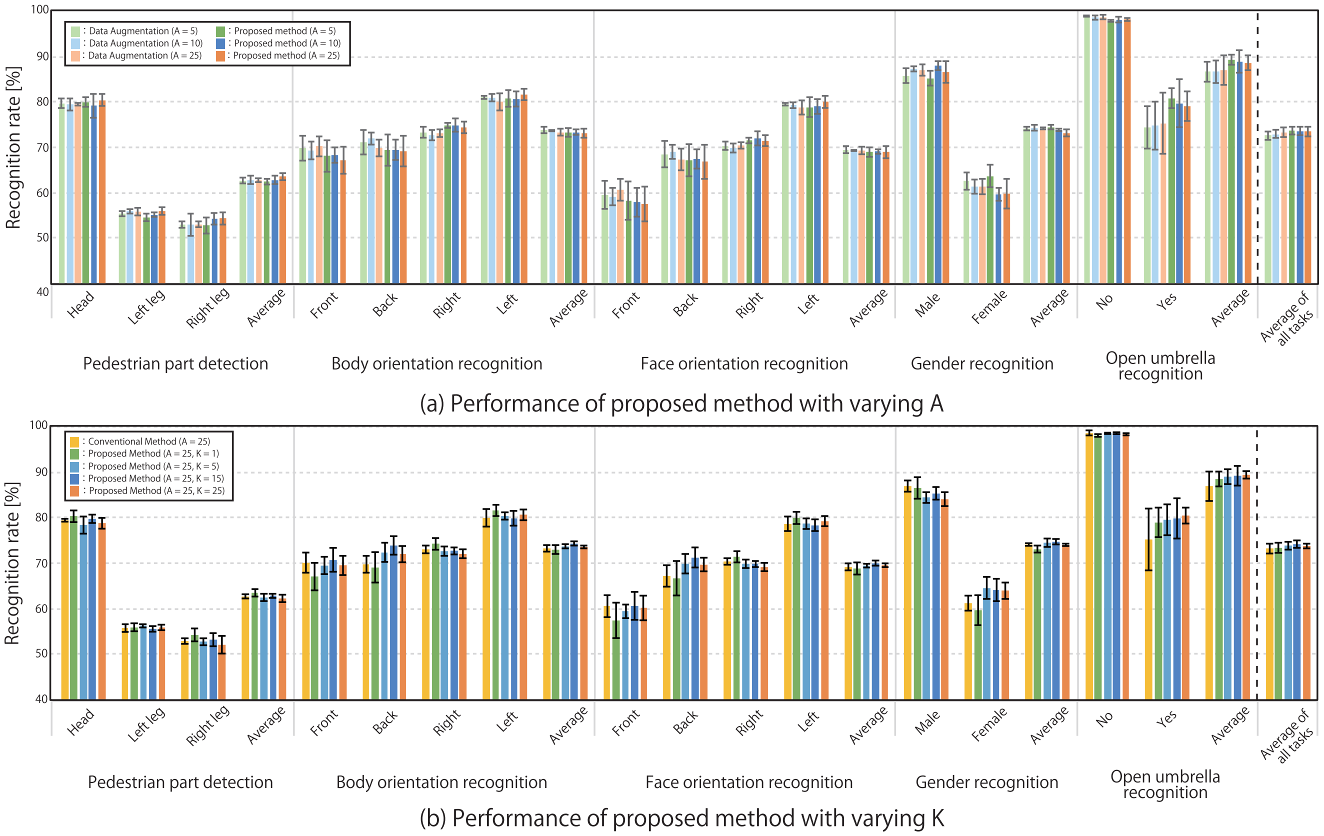
```
Data: Training dataset $\mathbf{D} = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^{N}$
// Assign rarity rate to training samples
for 0 to $N$ do
    $R_n = \sum_{t=1}^{T-1} (1 - p_{n,t}) \cdot \sqrt{1 - \frac{p_t^{min}}{p_t^{max}}} / T$
    $S_n = R_n \cdot A + 1$
    for 0 to $S_n$ do
        Increase number of training samples using data augmentation
    end for
end for
Data: Training dataset after data augmentation
    $\mathbf{D}' = \{(\mathbf{x}_{n'}, R_{n'}, \mathbf{y}_{n'})\}_{n'=1}^{N'}$
while not stopping criterion do
    // Assign rarity rate $R^B$ to mini-batch
    for 0 to $K$ do
        Create mini-batch by choosing $M$ training samples from training dataset $\mathbf{D}'$.
        $R^B_k = \sum_{m=1}^{M} R_m$
    end for
    Select mini-batch with median $R^B_k$ value.
    Calculate training error by inputting selected mini-batch into the CNN.
    Update weight filter and connection weight by backpropagation.
end while
```

3.4 Training Algorithm of the Proposed Method

Algorithm 1 is the training algorithm of the proposed method. First, it calculates the rarity rate R_n of the training sample in the dataset $\mathbf{D} = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^{N}$. It defines the number of augmentation samples S_n according to the rarity rate R_n of training sample n. Then, it applies data augmentation to the training sample. The augmented dataset $\mathbf{D}' = \{(\mathbf{x}_{n'}, R_{n'}, \mathbf{y}_{n'})\}_{n'=1}^{N'}$ is prepared. The CNN is trained by backpropagation using the mini-batch approach. It creates K mini-batch candidates and calculates their rarity rates R^B. The candidates are sorted by the mini-batch rarity rates. The mini-batch candidate with the median rarity rate is selected as the mini-batch to use for the training at this iteration. The training process of the heterogeneous learning CNN proceeds as described in Sect. 2.2.

4. Experiments

We evaluated our method for multiple tasks with datasets



(a) Performance of proposed method with varying A

(b) Performance of proposed method with varying K

Fig. 7 Performance of proposed method with varying parameters

that are unbalanced in the number of samples for each class. In the experiments, we investigated the best parameters for our method and compared the results with of other methods. The parameters of our method are the maximum number of augmentation samples A and the number of mini-batch candidates K. Changing the parameter A shows the effect of data augmentation based on the rarity rate. K shows the effect of selecting a mini-batch from a particular number of candidates. Once the optimal parameters were selected, we compared the performance of our method with that of 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.
- Evaluation of applying the proposed method for VGG16 [21].

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 divided these images into 45,581 training samples and 36,783 test samples. Each pedestrian image is labeled for five tasks: pedestrian part position, body orientation, face orientation, gender, and open umbrella recognition in Fig. 1. These ratios of training samples in each task are shown in Fig. 3. The class with the fewest samples is "open umbrella" in all of the tasks. The performance of each task except the separate pedestrian part detection is evaluated according to the recognition rate. For pedestrian part detection, we evaluate the localization error as a fraction of the head-to-toe 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-toe distance.

Table 1 Detail of our original CNN

Layer name	Detail	Size
Input	Gray scale image	1 x 64 x 128
conv.1	5 x 9 x 32	60 x 120 x 32
maxout 1	2	60 x 120 x 16
pool.1	2 x 2	30 x 60 x 16
conv.2	3 x 5 x 16	28 x 56 x 16
maxout 2	2	28 x 56 x 8
pool.2	2 x 2	14 x 28 x 8
conv.3	3 x 5 x 64	12 x 24 x 64
maxout 3	2	12 x 24 x 32
fc.1	sigmoid	3000
fc.2	sigmoid	2000
fc.3	sigmoid	1000
Output	heterogeneous	18

In this experiment, we used an original CNN that consists of three convolutional layers and three fully connected layers, as shown in Table 1. The total number of iterations to update the parameters is 500,000, and the training ratio η is set to 0.001. We evaluated each method five times to smooth

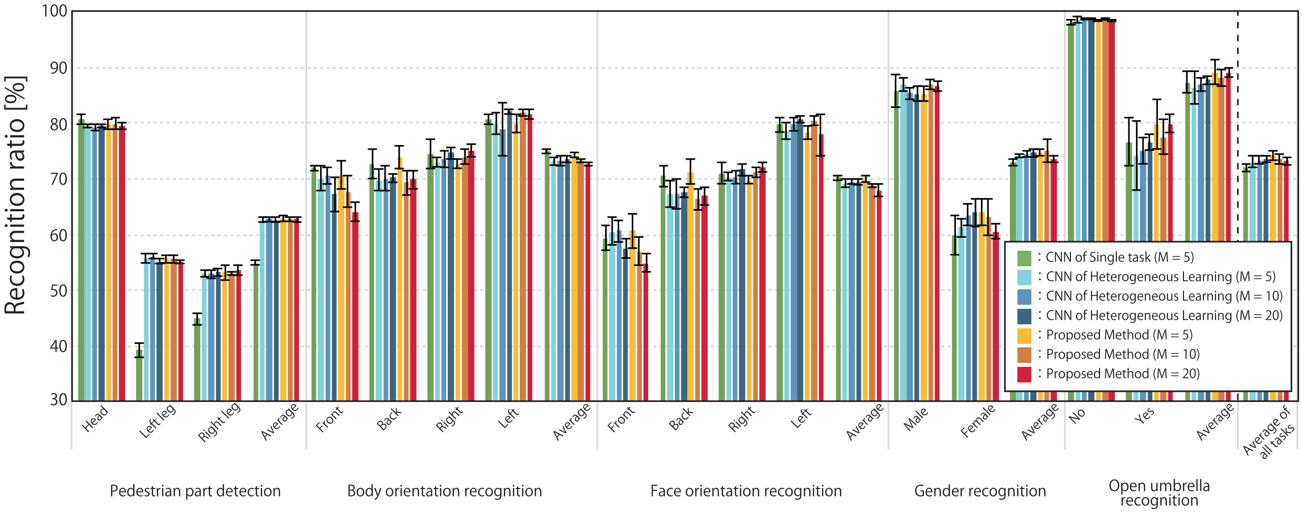


Fig. 9 Comparison between conventional methods and proposed method.

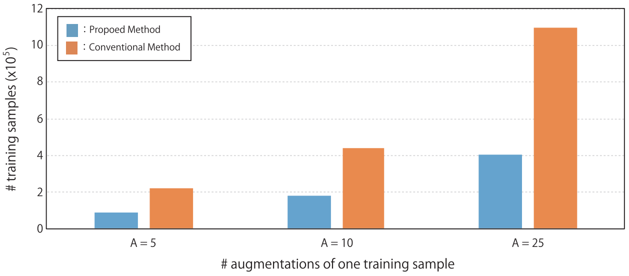


Fig. 8 Number of training samples used in conventional methods and proposed method

variability and calculated the standard deviation. In addition, we set the same initial parameters for all methods.

4.1 Performance for Varying Parameters

We evaluated the performance as we varied the maximum augmentation samples A and the number of mini-batch candidates K. Figure 7 (a), (b) show the recognition performance of the proposed method with varying A and K, respectively. The accuracy is the average from five runs, and the error bar denotes the standard deviation. Figure 7 (a) compares the recognition rates when the parameter A is {5, 15, 25}. We also evaluated the effect of varying parameter A with and without the rarity rate. When we set the A to 5, we obtained the best performance and this improves the accuracy of the rare class "open umbrella" by about 5.5%. However, the accuracy of the classes "body orientation is front" and "face orientation is front" and "female gender" are lower than with the conventional method. These classes have the fewest training samples in the body orientation task, face orientation task, and gender task, respectively, because it is assumed that the deflection between classes in "open umbrella" is about 10% more than that in other tasks. Figure 8 shows the number of training samples used in the training. When the number of training samples increases to 25, the best performance is achieved with the conventional method. In contrast, when we set the maximum augmentation sample parameter A to 5, the best performance is achieved with the proposed method. As shown in Fig. 8, the conventional method uses about 1.1 million training samples. The conventional method are evenly augmented all training samples using data augmentation at {5, 10, 25} times. However, the proposed method uses about 0.4 million training samples. This means that the proposed method reduces the number of training samples by more than half by suppressing the unnecessary common samples that would otherwise be increased by data augmentation. As shown in Fig. 7 (a) and Fig. 8, in the proposed method, it is easy to choose the rare samples by suppressing the augmented common samples.

Figure 7 (b) shows a comparison of the recognition rates as the parameter K varies between {1, 5, 15, 25}. In this experiment, we evaluated by fixing $A = 25$ in Fig. 7 (b). For this reason, the performance when $A = 25$ is better than the performance with other values of parameter A when changing the parameter K. When K is set to 15, the best performance of about 74.18% was achieved. In addition, it is possible to suppress the standard deviation by half by increasing the number of mini-batch candidates K. While the performance of the classes "body orientation is front", "face orientation is front", and "female gender" are lower than with the conventional method, as shown in Fig. 7 (a), the performance of these classes is improved to be equal to or higher than the conventional method by changing the parameter K to 15. It is easy to input a mini-batch that has moderate "body orientation is front", "face orientation is front", and "female pedestrian" to a CNN by increasing the number of candidate mini-batches K.

4.2 Performance Comparison for Multi-Task Classification

In Fig. 9, we compared the performance of our proposed

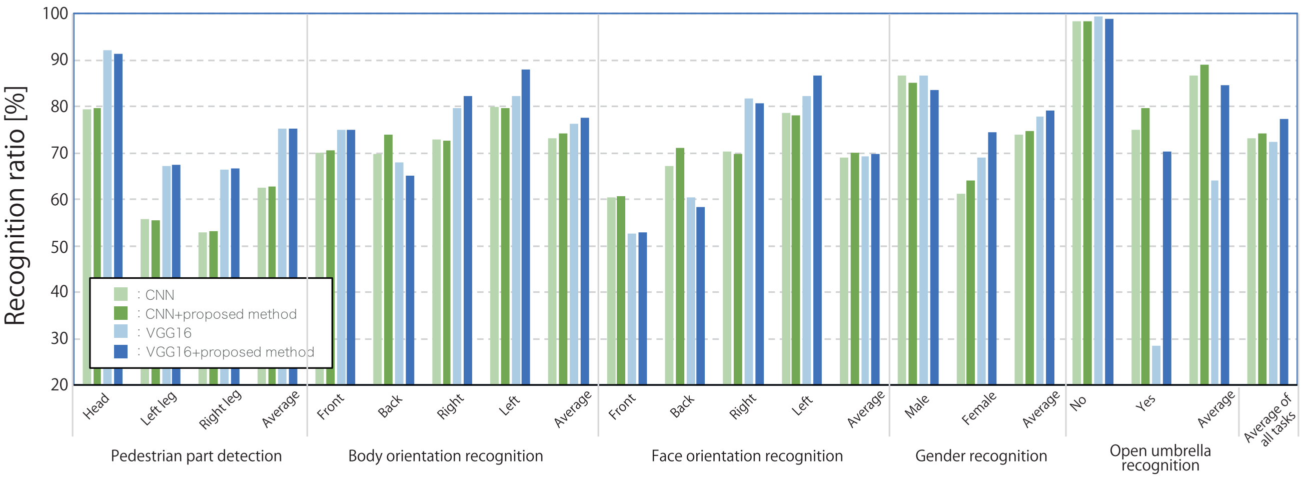


Fig. 10 Evaluation performance of VGG16

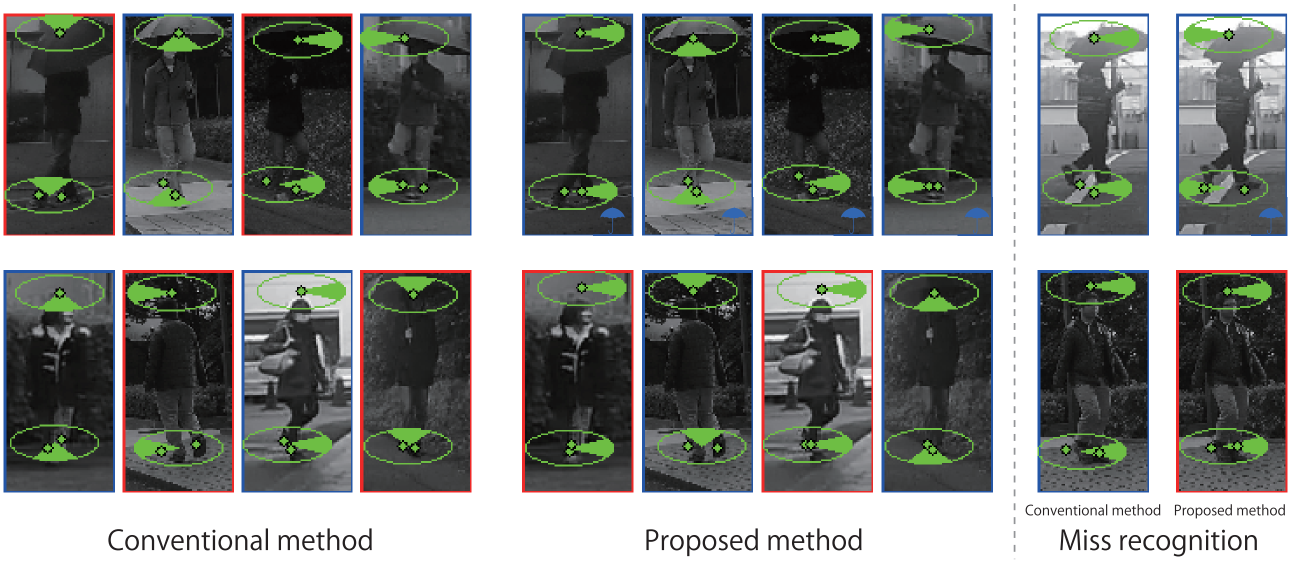


Fig. 11 Comparison of example pedestrian attribute recognition using conventional method and proposed method

method with single task learning and heterogeneous learning. We show the recognition rates as the parameter M varies within {5, 10, 20}.

Figure 9 shows that the CNN with heterogeneous learning is superior to 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 compared with the single task approach. The body orientation attribute is supported by part detection, especially the detection of the legs, because the part position and body orientation recognition can be performed simultaneously. The proposed method improves the performance for rare classes by applying the rarity rate to heterogeneous learning. When we evaluated the various values of 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.

Figure 10 shows the performance of applying the proposed method when applied for VGG16. In this experiments, parameters A and K of the proposed method are 25 and 15. When we compare the performance of CNN, VGG16 and proposed method, VGG16 with proposed method is the best performance in an accuracy of all tasks. In the classes of low training sample, the proposed method is better accuracy than conventional VGG16 except "body orientation is back" and "face orientation is back". In particular, "open umbrella" performance improves by more than 50% for VGG16 with the proposed method. However, when we compare the performance of CNN and VGG16 with the proposed method, do not improve performance of "body orientation is back", "face orientation is front", "face orientation is back", "gender is male" and "open umbrella".

In Fig. 11, we show the recognition results. The circles at the head and feet denote the orientation of the face and body, respectively. The green points are the detected part positions at the head and legs. The color of the bounding box indicates the gender (blue is male and red is female). In addition, when an open umbrella is detected, an umbrella icon is shown at the bottom-right. As shown in Fig. 11, 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.

5. Conclusion

In this paper, we proposed a method to improve the performance of heterogeneous learning of multi-task pedestrian attribute recognition for classes with few samples. We assigned a rarity rate to each training sample to determine the number of augmentations such that rare samples are augmented more. In addition, we created multiple mini-batch candidates from the augmented datasets and the candidate that has an appropriate balance 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 samples. By introducing the rarity rate during data augmentation, the proposed method reduces the number of training samples by half.

Acknowledgments

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

References

[1] D. Gerónimo, A.M. López, A.D. Sappa, and T. Graf, "Survey of Pedestrian Detection for Advanced Driver Assistance Systems," IEEE Trans. Pattern Anal. Mach. Intell., 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, pp.886–893, 2005.

[3] P. Felzenszwalb, D. McAllester, and D. Ramaman, "A Discriminatively Trained, Multi scale, Deformable Part Model," Computer Vision and Pattern Recognition, pp.1–8, 2008.

[4] X. Wang, T.X. Han, and S. Yan, "An HOG-LBP Human Detection with Partial Occlusion," International Conference on Computer Vision, pp.32–39, 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, pp.1801–1808, 2011.

[6] J. Marin, D. Vazquez, A.M. Lopez, J. Amores, and B. Leibe, "Random Forests of Local Experts for Pedestrian Detection," International Conference on Computer Vision, pp.2592–2599, 2013.

[7] H. Fukui, T. Yamashita, Y. Yamauchi, H. Fujiyoshi, and H. Murase, "Pedestrian Detection Based on Deep Convolutional Neural Network with Ensemble Inference Networks," 2015 IEEE Intelligent Vehicles Symposium (IV), pp.223–228, 2015.

[8] Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian Detection aided by Deep Learning Semantic Tasks," Computer Vision and Pattern Recognition, pp.5079–5087, 2015.

[9] J. Hosang, M. Omran, R. Benenson, and B. Schiele, "Taking a Deeper Look at Pedestrian," Computer Vision and Pattern Recognition, pp.4073–4082, 2015.

[10] H. Brodsky, and A.S. Hakkert, "Risk of a road accident in rainy weather," Accident Analysis & Prevention, vol.20, no.3, pp.161–176, 1988.

[11] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex Multi-task Feature Learning," Mach. Learn., vol.73, no.3, pp.243–272, 2008.

[12] X. Yang, S. Kim, and F.P. Xing, "Heterogeneous Multi-task Learning with Sparsity Constrain," Advances in Neural Information Processing Systems 22, pp.2151–2159, 2009.

[13] 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, pp.488–495, 2014.

[14] Z. Zhang, P. Luo, C.C. Loy, and X. Tang, "Facial Landmark Detection by Deep Multi-task Learning," Computer Vision – ECCV 2014, Lecture Notes in Computer Science, vol.8694, pp.94–108, Springer International Publishing, Cham, 2014.

[15] F.J. Pulgar, A.J. Rivera, F. Charte, and M.J.D. Jesus, "On the Impact of Imbalanced Data in Convolutional Neutral Networks Performance," Hybrid Artificial Intelligent Systems, Lecture Notes in Computer Science, vol.10334, pp.220–232, Springer International Publishing, Cham, 2017.

[16] C. Huang, Y. Li, C.C. Loy, and X. Tang, "Learning Deep Representation for Imbalanced Classification," IEEE Computer Vision and Pattern Recognition, pp.5375–5384, 2016.

[17] J. Zhu, S. Liao, D. Yi, Z. Lei, and S.Z. Li, "Multi-label CNN Based Pedestrian Attribute Learning for Soft Biometrics," International Conference on Biometrics, pp.535–540, 2015.

[18] D. Li, X. Chen, and K. Huang, "Multi-attribute Learning for Pedestrian Attribute Recognition in Surveillance Scenarios," International Association of Pattern Recognition on Asian Conference on Pattern Recognition, pp.111–115, 2015.

[19] E. Bekele, C. Narber, and W. Lawson, "Multi-attribute Residual Network (MAResNet) for Soft-biometrics Recognition in Surveillance Scenarios," IEEE International Conference on Automatic Face and Gesture Recognition, pp.386–393, 2017.

[20] 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.

[21] K. Simonyan, and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," International Conference on Learning Representations, 2015.

[22] 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, pp.4660–4668,2015.

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

[24] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian Detection: An Evaluation of the State of the Art," IEEE Trans. Pattern Anal. Mach. Intell., vol.34, no.4, pp.743–761, 2012.

[25] D.E. Rumelhart, G.E. Hinton, and R.J. Williams, "Learning representations by back-propagating errors," Nature., vol.323, no.6088, pp.533–536, 1986.

Hiroshi Fukui received his BEng and MEng from Department of Computer Science, Chubu University in 2016. His research interests include computer vision and machine learning.

Takayoshi Yamashita received his Ph.D degree from Department of Computer Science, Chubu University, Japan in 2011. He worked in OMRON Corporation from 2002 to 2014. He is a lecturer of the Department of Computer Science, Chubu University, Japan since 2014. He current research interests include object detection, object tracking, human activity understanding, pattern recognition and machine learning. He is a member of the IEEE, the IEICE and the IPSJ.

Yuji Yamauchi received the Ph.D. from Department of Computer Science, Chubu University in 2012. From 2012 to 2014 he was a Ph.D. fellow at the Chubu University. In 2011, he was a visiting student at Robotics Institute, Carnegie Mellon University. He was a Fellowship of the Japan Society for the Promotion of Science from 2010 to 2012. His research interests include computer vision and pattern recognition. He is a member of the IEEE, the IEICE and the IPSJ.

Hironobu Fujiyoshi received his Ph.D. in Electrical Engineering from Chubu University, Japan, in 1997. From 1997 to 2000 he was a post-doctoral fellow at the Robotics Institute of Carnegie Mellon University, Pittsburgh, PA, USA, working on the DARPA Video Surveillance and Monitoring (VSAM) effort and the humanoid vision project for the HONDA Humanoid Robot. He is now a professor of the Department of Computer Science, Chubu University, Japan. From 2005 to 2006, he was a visiting researcher at Robotics Institute, Carnegie Mellon University. His research interests include computer vision, video understanding and pattern recognition. He is a member of the IEEE, the IEICE, the IPSJ, and the IEEE.

Hiroshi Murase received the BEng, MEng, and PhD degrees in electrical engineering from Nagoya University, Japan. In 1980 he joined the Nippon Telegraph and Telephone Corporation (NTT). From 1992 to 1993 he was a visiting research scientist at Columbia University, New York. From 2003 he is a professor of Nagoya University, Japan. He was awarded the IEICEJ Shinohara Award in 1986, the Telecom System Award in 1992, the IEEE CVPR (Conference on Computer Vision and Pattern Recognition) Best Paper Award in 1994, the IPSJ Yamashita Award in 1995, the IEEE ICRA (International Conference on Robotics and Automation) Best Video Award in 1996, the Takayanagi Memorial Award in 2001, the IEICEJ Achievement Award in 2002, and the Ministry Award from the Ministry of Education, Culture, Sports, Science and Technology in 2003. Dr. Murase is IEEE Fellow, IEICEJ Fellow, and a member of the Information Processing Society of Japan.