Facial Image Analysis by CNN with Weighted Heterogeneous Learning

Hiroshsi Fukui, Takayoshi Yamashita, Yuu Kato, Ryo Matsui, Yuji Yamauchi, Hironobu Fujiyoshi
College of Engineering, Chubu university
1200 Matsumoto-cho, Kasugai City, Aichi Prefecture, Japan
{fhiro@vision., yamashita@, yuu@isc., hf@} cs.chubu.ac.jp

***Abstract*—Recognition of facial attributes such as facial point, gender, and age have been used in marketing strategies and client services on social networks. In general, to recognize these attributes, it requires independent handcraft features and classifiers for each task. Heterogeneous learning is able to train a single classifier to perform multiple tasks. This learning method simultaneously train regression and recognition tasks, thereby reducing both training and testing time. However, differences between training error negatively affect the training process in specific tasks. To address this problem. we propose weighted heterogeneous learning which has weighed error function for a deep convolutional neural network. Our method outperformed the conventional method in terms of facial attribute recognition, especially for regression tasks such as facial point detection, age estimation, and smile ratio estimation.**

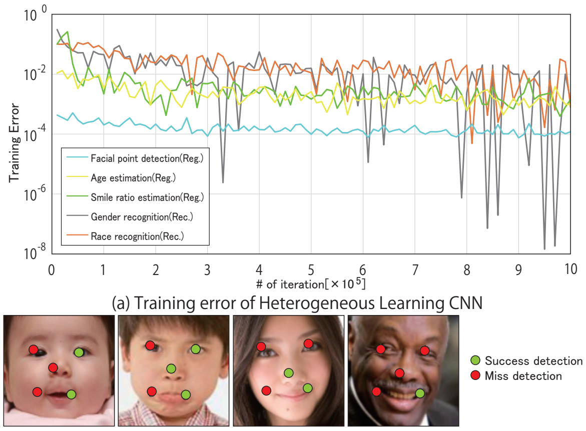
***Index Terms*—Deep Convolutional Neural Network, Heterogeneous learning, Facial image analysis**

I. Introduction

Recognition of facial attributes such as facial point, gender, and age has been used in marketing strategies and social networking services. Marketing strategies recommend the goods, that are supposed to matches the needs of potential clients. Various social networking services based on facial recognition techniques have recently been developed that can estimate age from a facial image with a high accuracy.

To recognize multiple face attributes, it is necessary to train classifier for each task, such as facial point detection, gender recognition, and age estimation. Active appearance model (AAM) [1] and conditional regression forest (CRF) [2] are common approaches for facial point detection. Additionally, age estimation and gender recognition are classified by a support vector machine (SVM) or decision tree using facial point or a local binary pattern (LBP) features [3][4]. With the increasing of deep learning, the deep convolutional neural network (CNN) [5] has become a common classifier for facial point detection [6][7][8], age estimation [9][10], and gender recognition [11][12].

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. However, training errors of each task is quite wide. This difference of training error is occured by difference between label range of regression task and recognition task. The label range of regression tasks is a continuous value from 0 to 1, whereas the label range for classification tasks is a discrete value of 0 or 1. Consequently, facial point detection performance suffers, as shown in Fig. 1(b). Therefore, differences between training errors negatively affect the training process for heterogeneous learning for a CNN.



(a) Training error of Heterogeneous Learning CNN

(b) Examples of visualizing facial point detection

Fig. 1. Training error and example results by heterogeneous learning

In this paper, we propose weighted heterogeneous learning for a CNN. First, we select a basis task from all tasks. Additionally, other tasks are defined subtasks. After defining basis task and subtasks, we weight the error function for the subtasks. Our method suppresses the training error and dispersion training errors by weighting the cost function for the subtasks. Weighted heterogeneous learning for a CNN improves the recognition performance by stable training.

II. Facial image analysis using Heterogeneous Learning CNN

First, we describe the related publications in these categories and then further discuss problems with existing heterogeneous learning for a CNN method as applied to facial image analysis.

A. Related work

Marketing strategies and social networking services have used facial attribute information, such as facial point, gender, and age. In particular, facial point have been used as features for estimating age, gender, and facial expressions. AAM [1] is a common approach for facial point detection. AAM detects optimal facial point by changing face model parameters iteratively. AAM can detect facial point to a high accuracy

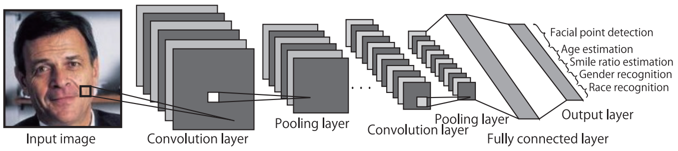


Fig. 2. Heterogeneous learning for a CNN

for use in training facial images. However, it is difficult for an unknown testing sample to detect facial point. The CRF proposed by Dantone et al. detects facial point using regression forests for each face pose [2]. CRF consists of two stages: the first estimates the facial pose, and the second regresses the facial point using regression forests. Age estimation and gender recognition are classified by a SVM or decision tree using facial point or LBP features [3][4]. CNN has also become a common classifier for facial point detection [6][7][8], age estimation [9][10], and gender recognition [11][12].

Performing recognition or estimation for multiple tasks requires the construction of classifiers corresponding to each task. However, 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, which 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. Heterogeneous learning can estimate and recognize multiple facial attribute with high accuracy by combining CNN [14][15]. Zhang et al. proposed a method to perform multiple tasks such as facial point detection, gender classification, face orientation estimation, and glasses detection [14]. While the method estimated multiple tasks, its main purpose was to improve the performance of the primary task, such as facial point detection. It thus assigned weighted error functions to each task. When the error decreased sufficiently, the training of the task was terminated earlier to avoid over-fitting to a specific task.

B. Heterogeneous Learning

Figure 2 shows the structure of a heterogeneous learning for a CNN. First, M training samples are chosen randomly to form a mini-batch. We used mini-batch training when updating CNN parameters. During mini-batch training, the error E is calculated and backpropagated to update the parameters θ of the network. For each backpropagation [16] iteration, the samples in the mini-batch are selected randomly from the dataset. When the CNN is trained using heterogeneous learning, the recognition and regression tasks are combined in a single network and each task has an independent error function. The mean squared error in Eq.(1) and the cross entropy in Eq.(2) are employed as the error functions of the regression and recognition tasks, respectively.

$$E_t^{Regression} = \frac{1}{M}\sum_{m=1}^{M} \parallel \mathbf{y} - \mathbf{o} \parallel_2^2 \qquad (1)$$

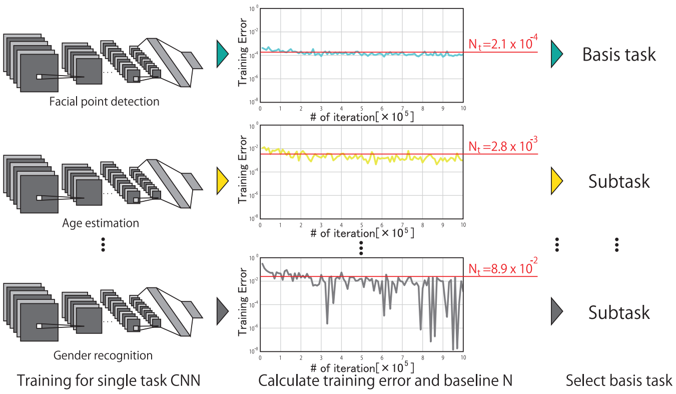


Fig. 3. Selection basis task

$$E_t^{Recognition} = \frac{1}{M}\sum_{m=1}^{M} -\mathbf{y}\log\mathbf{o} \qquad (2)$$

The errors E_t of the sample m for all tasks $\{t|1,\ldots,1\}$ are accumulated and propagated once per iteration.

C. CNN based on heterogeneous learning

Figure 1(a) shows the training errors of five tasks using heterogeneous learning for a CNN. There are differences between the training errors for all tasks.

The differences between training errors occur because of the error function for regression tasks and recognition tasks. There are noticeable differences between the error ranges of the mean squared error function and the cross entropy error function. The mean squared error ranges from 0 to 1, and the cross entropy error ranges from 0 to infinity. Thus, we integrate the error range from 0 to 1 by exchanging the cross entropy error function for the mean squared error function for recognition tasks.

However, the differences between training error occur if integration errors range from 0 to 1 by exchanging the cross entropy error function for the mean squared error function for recognition tasks. The label range of regression tasks is a continuous value from 0 to 1, whereas the label range of recognition tasks is a discrete value of 0 or 1. Thus, recognition tasks develop more differences between the training errors than regression tasks. These causes negatively affect heterogeneous learning during the training process. Thus, facial point detection performance suffers due to the lowest training error, as shown in Fig. 1(b).

III. Proposed Method

Conventional heterogeneous learning calculates the training error under Eq. (1) evenly. Hence, differences between training errors occur because of differences between label ranges for regression tasks and recognition tasks. The proposed method stabilizes the training error by weighting each task and improves the heterogeneous learning performance.

First, we obtain training error by training CNN of a single task for each task. Unlike in training error of heterogeneous learning, the training error of a single task CNN is not

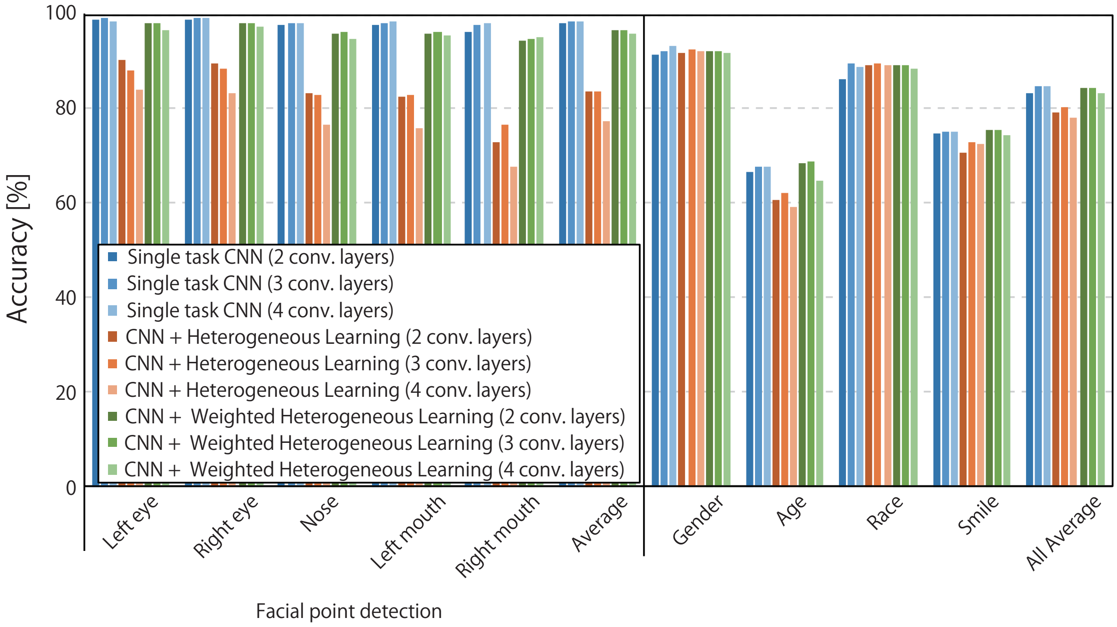


Fig. 4. Comparison of the proposed method and other method

interference training error between other tasks. Therefore, we will be able to obtain a stable basis value using training error of single task CNN, when computing basis values for each task.

Basis value N_t for each task that gave weight to error functions are computed by using training error of single task CNN, as shown in Fig. 3. These basis value N_t are used the normal distribution of the training error in Eq. (3). If it reflects training error for normal distribution, the normal distribution of the training error for task t connotes 99.7% in the interval that sums the average μ and 3-fold vertical 3σ. The other interval is the dispersion of training error, and we can calculate the basis value N_t that is negatively affected by ignoring the interval.

$$N_t = \mu + 3\sigma \tag{3}$$

We select a basis task from the lowest basis value N_t. Thus, the facial point detection task is the basis task and the other tasks are subtasks. After selecting the basis task and subtasks, we calculate the weight w_t for each subtask. The basis value N_f of the facial point detection task and basis value N_t of the other tasks are used in Eq. (4).

$$w_t = \frac{N_f}{N_t} \tag{4}$$

We give weight to error function for each subtask, as shown in Eq. (5). The first term in Eq. (5) is an error function of the main task. The second term in Eq. (5) is an error function of subtasks.

$$E = \frac{1}{M}\sum_{m=1}^{M}\left(||\boldsymbol{y}_{f,m} - \boldsymbol{o}_{f,m}||_2^2 + \sum_{t=1,t\neq f}^{T-1} w_t||\boldsymbol{y}_{t,m} - \boldsymbol{o}_{t,m}||_2^2\right) \tag{5}$$

Note that $\boldsymbol{y}_{f,m}$ and $\boldsymbol{o}_{f,m}$ are the label range and output of the facial point detection task, respectively. Additionally, the weight w_t is constant for each iteration.

TABLE I
PARAMETERS OF HETEROGENEOUS LEARNING CNN STRUCTURE

	1st	2nd	3rd	4th	5th
2 conv. layers	C(16, 9 × 9)	C(32, 9 × 9)	F(200)		
3 conv. layers	C(16, 9 × 9)	C(32, 9 × 9)	C(64, 9 × 9)	F(200)	
4 conv. layers	C(16, 9 × 9)	C(32, 9 × 9)	C(64, 9 × 9)	C(128, 9 × 9)	F(200)

IV. Experiments

We evaluate the proposed method by comparing its performance with those of the CNN for a single task and conventional heterogeneous learning. In these experiments, we perform facial point detection, gender recognition, race recognition, age estimation and smile ratio estimation. For facial point detection, five facial points such as left eye, right eye, nose, left mouth, and right mouth are detected by regression estimate. We employ three type CNN, as shown tab. I. In tab. I, C(·) and F(·) mean convolution layer and means fully connection layer, respectively. Note that these CNNs apply max pooling at 1st layer and 2nd layer, and applying maxout at all convolution layers. Smile ratio estimation is identified as regression of the value between 0 and 99. Note that, smile ratio label is the average of some smile ratios that some people are given as labels. Age label is identified as regression of the value between 0 and 100. Race recognition is identified as Asian, White, or Black.

The total number of iterations to update the parameters is 1,000,000, the training coefficient η is set to 0.001, and the mini-batch size is 10. The comparison dataset consists of 53,663 facial images that were captured by aggregating face images from the Web. However, almost no published dataset has been given any facial attribute labels, because we created a facial attribute dataset that has been given five facial attribute. Note that we performed 5-fold cross validation: 42,663 images used for training, and 11,000 images for testing. The image size is 100×100 pixel with grayscale. The evaluation method of facial point detection is the same as that of Dantone et

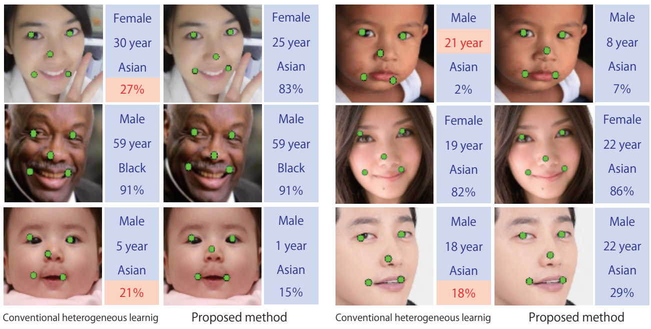


Fig. 5. Comparison of examples of facial image analysis

al. [2]. In age and smile ratio estimation, we judge estimation to be successful if the difference between output and label is connoted by the threshold, which are ±5 years and 10%.

A. *Comparison of performance for the proposed method*

Figure 4 shows the performances of single task learning, conventional heterogeneous learning and the proposed method. The single task learning obtains better accuracy than conventional heterogeneous learning in the regression tasks, especially for the facial point detection task. In comparison the proposed method with conventional heterogeneous learning, the proposed method improves the all average accuracy by approximately 5% and the accuracy of the facial point detection task by approximately 14%. In comparison with number of convolution layer at CNN, improved performance by adding convolution layer. However, proposed method is best performance when using "3 conv. layers" CNN. This means that proposed method is able to extract the features stability when using small CNN.

Figure 5 shows an example of facial image analysis using conventional heterogeneous learning and the proposed method. The first and third columns show result of examples of conventional heterogeneous learning, and the second and fourth columns show results of the proposed method. and the text on their right is results of subtasks such as gender, age, race, and smile ratio. The green points are facial points detected by conventional heterogeneous learning or proposed method. The red text is inaccurate recognition or estimation. As shown in Fig. 5, we observe that the proposed method is robust to faces with large pose variation, lighting, and severe occlusion. Additionally, the processing time of our method is approximately 22ms to analyze one image on an Intel Core i7-4790 (3.4GHz) with 8GB of memory, and the processing time is approximately 1.8 ms to analyze one image on GeForce GTX980.

B. *Comparison of training errors*

Figure 6 shows the training error for each task for the proposed method at "3 conv. layers". The training error for conventional heterogeneous learning is different for each task, and the training error varies suddenly for the recognition task, as shown in Fig. 1(a). Additionally, training errors of the proposed method for each task are lower overall than those of conventional heterogeneous learning. The proposed method has a unified training error for each task, and suppresses the dispersion training error variation. To achieve this result, the proposed method is stably trained by weighting the error function.

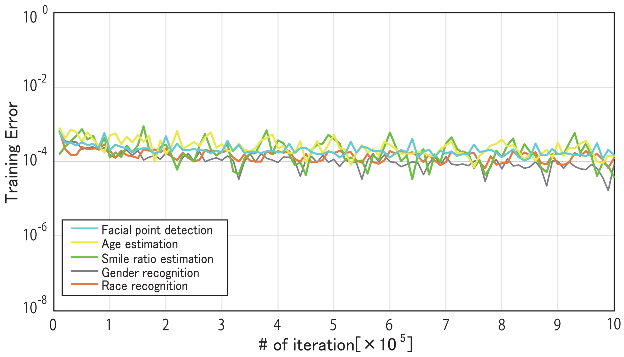


Fig. 6. Comparison of the raining error for proposed method

V. Conclusion

We proposed a method to improve the performance of heterogeneous learning for facial image analysis. As a result, the proposed method improved performance by approximately 5% and the accuracy of the facial point detection task by approximately 14%.

References

[1] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. ECCV, 1998.

[2] M. Dantone, J. Gall, G. Fanelli, and L. V. Gool. Real-time facial feature detection using conditional regression forests. CVPR, 2012.

[3] W. Jun, Z. Yi, J. M. Zurada, B. L. Lu, H. Yin. Multi-view Gender Classification Using Local Binary Patterns and Support Vector Machines. In Third International Symposium on Neural Networks, 2006.

[4] G. Guodong, F. Yun, R. D. Charles, and S. H. Thomas. Image-Based Human Age Estimation by Manifold Learning and Locally Adjusted Robust Regression. IEEE Transactions on Image Processing, Vol.17, pp.1178-1188, 2008.

[5] A. Krizhevsky, S. Ilva, G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Network. NIPS, pp.1097-1105, 2012.

[6] Y. Sun, X. Wang, and X. Tang. Deep Convolutional Network Cascade for Facial Point Detection. CVPR, 2013.

[7] M. Kimura, T. Yamashita, Y. Yamauchi, and H. Fujiyoshi. Facial point detection based on a convolutional neural network with optimal mini-batch procedure. ICIP, 2015.

[8] A. Jourabloo, and X. Liu. Large-pose Face Alignment via CNN-based Dense 3D Model Fitting. CVPR, 2016.

[9] Y. Zhu, L. Yan, M. Guowang, and G. Guodong. A Study on Apparent Age Estimation. ICCV Workshops, 2015.

[10] K.Zhanghui, H. Chen, Z. Wei. Deeply Learned Rich Coding for Cross-Dataset Facial Age Estimation.ICCV Workshops, pp. 96-101, 2015.

[11] A. Grigory, B. Sid-Ahmed, and D. Jean-Luc. Minimalistic CNN-based ensemble model for gender prediction from face images. Pattern Recognition Letters, Vol. 70, pp. 59-65, 2015.

[12] G. Levi, and T. Hassner. Age and Gender Classification Using Convolutional Neural Networks. CVPR, 2015.

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

[14] Z. Zhang, P. Luo, C. C. Loy, X. Tang. Facial Landmark Detection by Deep Multi-task Learning. ECCV, 2014.

[15] R. Ranjan, M. P. Vishalx, and R. Chellappa. Facial Landmark Detection by Deep Multi-task Learning. arXiv preprint arXiv:1603.01249, 2016.

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