Multiple Facial Attributes Estimation based on Weighted Heterogeneous Learning

H.Fukui* T.Yamashita* Y.Kato* R.Matsui*
T. Ogata** Y.Yamauchi* H.Fujiyoshi*

*Chubu University
1200, Matuoto-cho, Kasugai,
Aichi, Japan

**Abeja Inc.
4-1-20, Toranomon, Minato-ku,
Tokyo, Japan

Abstract. To estimate multiple face attributes, independent classifier for each attribute are trained such as facial point detection, gender recognition, and age estimation in the conventional approach. It is inefficient because the computational cost of training and testing increases with the number of tasks. To address this problem, heterogeneous learning is able to train a single classifier to perform multiple tasks. Heterogeneous learning is simultaneously train regression and recognition tasks, thereby reducing both training and testing time. However, it is difficult to obtain equivalent performance for set of single task classifiers due to variance of training error of each task. In this paper, we propose weighted heterogeneous learning of a convolutional neural network with a weighted error function. 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.

1 Introduction

Facial attributes estimation such as facial point, gender, and age has been used in marketing strategies and social networking services. Marketing strategies recommend specific items that are matched to the client requirement. 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 estimate multiple face attributes, independent classifier for each attributes are trained 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. Age estimation and gender recognition are classified by a support vector machine (SVM) or a decision tree using facial points or local binary pattern (LBP) features [3][4][5]. With the increase of deep learning, the deep convolutional neural network (CNN) [6] has become a common classifier for facial point detection [7][8][9][10][11][12], age estimation [13][14][15][16], and gender recognition [17][18][19]. The conventional approach must prepare multiple classifiers for each task. This approach is inefficient because the computational cost of training and testing increases with the number of tasks. To address this problem, heterogeneous learning [20] trains

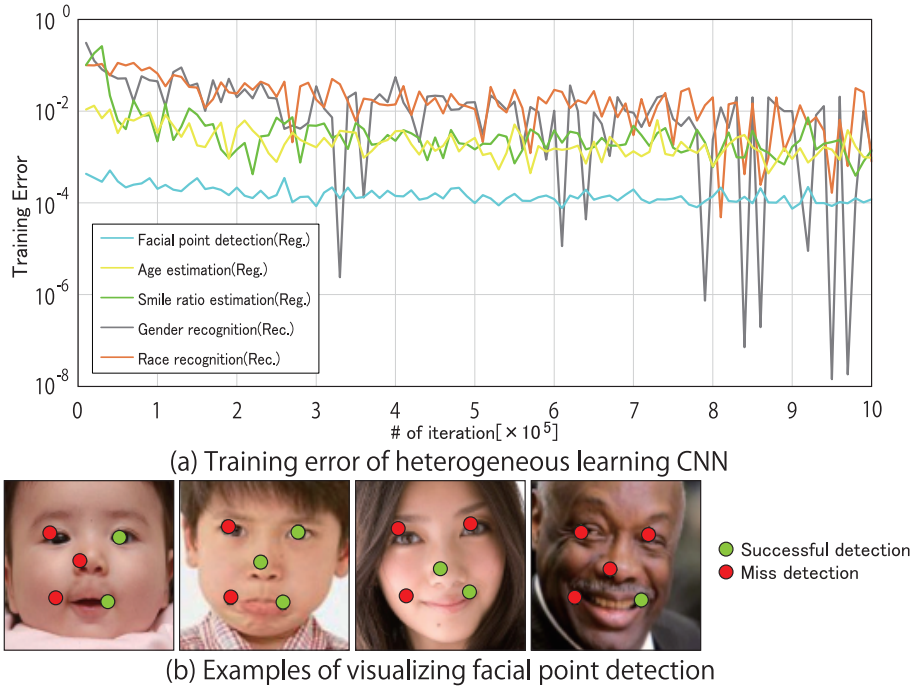

Fig. 1. Training error and example results by heterogeneous learning for a CNN

a single classifier to perform multiple recognition tasks. A CNN trained using heterogeneous learning has output units that correspond to each task. Thus, a single network classifies multiple tasks simultaneously and the computational cost does not vary with the number of tasks. In this paper, we use a heterogeneous learning CNN for facial point detection, gender recognition, age estimation, race recognition, and smile rate estimation.

Conventional heterogeneous learning has used the mean squared error function for regression tasks and the cross entropy error function for recognition tasks during the training process. The error ranges of mean squared error function and cross entropy error function are noticeably different. Therefore, we integrate the error range from 0 to 1 by exchanging cross entropy error function for mean squared error function for recognition tasks. However, if we integrate the training error functions, difference of training error is occured, as shown in Fig. 1(a). This difference of training error is occured by difference between label value of regression task and recognition task. The label value of regression tasks is a continuous value from 0 to 1, whereas the label value 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.

In this paper, we propose weighted heterogeneous learning for a CNN. First, we select a basis task from all tasks. Additionally, we define subtasks, not includ-

ing the basis task. 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 when introducing the proposed method.

2 Facial image analysis using heterogeneous learning CNN

We categorize related work into facial image analysis and heterogeneous learning. 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.

2.1 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 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][5]. CNN has also become a common classifier for facial point detection [7][8][9][10][11][12], age estimation [13][14][15][16], and gender recognition [17][18][19].

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 [21][22][23][24][25]. Zhang et al. proposed a method to perform multiple tasks such as facial point detection, gender classification, face orientation estimation, and glasses detection [21]. 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.

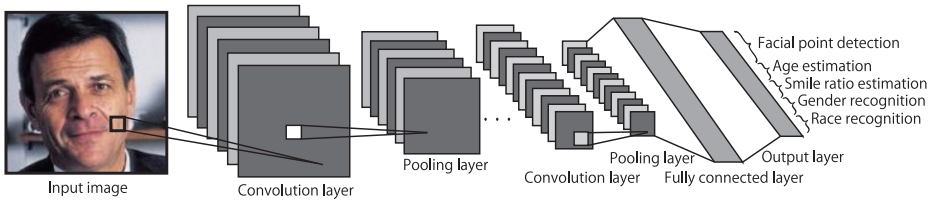

Fig. 2. Heterogeneous learning for a CNN

2.2 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[26] 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 recognition and regression tasks, respectively.

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

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

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

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

The parameters $\boldsymbol{\theta}$ of the CNN are updated using the differential of the accumulated error with the training coefficient η.

2.3 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

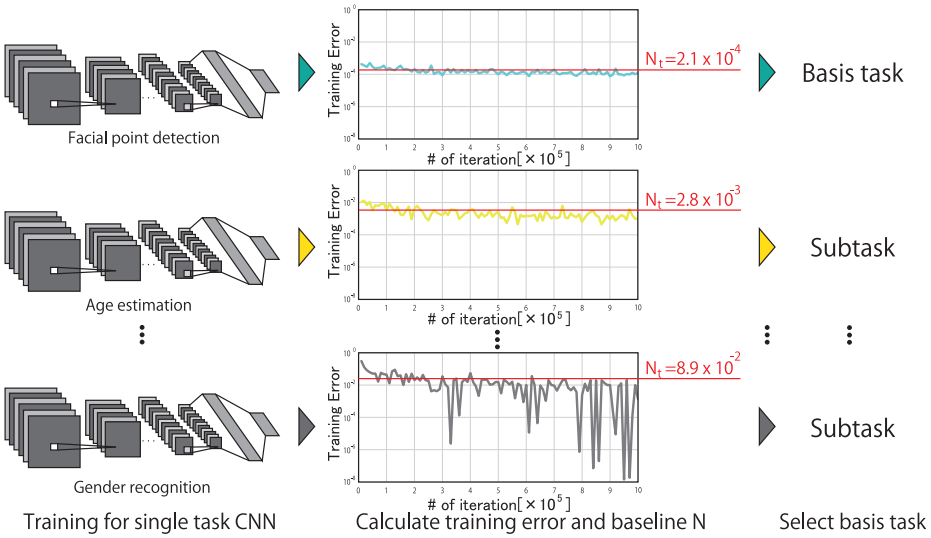

Fig. 3. Selection basis task

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 value of regression tasks is a continuous value from 0 to 1, whereas the label value 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).

3 Proposed method

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

3.1 Training of Single task CNN

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 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. In this paper, this CNN training is repeated until the training criterion condition is satisfied.

3.2 Computing the weights of error functions

We compute basis value N_t for each task that gave weight to training error functions using training error of single task CNN, as shown in Fig. 3. However, these training errors are varied by each iteration. Therefore, we calculate the basis value N_t using the normal distribution of the training error for each task. 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. Thus, we use the basis value N_t that sums the average μ and 3-fold vertical 3σ.

$$N_t = \mu + 3\sigma \tag{4}$$

After calculating the basis value, we select a basis task. In this paper, we select the basis task for 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).

$$w_t = \frac{N_f}{N_t} \tag{5}$$

3.3 Training of weighted heterogeneous learning

We give weight to error function for each subtask, as shown in Eq. (6). The first term in Eq. (6) is an error function of the main task. The second term in Eq. (6) is an error function of subtasks.

$$E = \frac{1}{M}\sum_{m=1}^{M}\left(||\boldsymbol{y}_{f,m} - \boldsymbol{o}_{f,m}||_2^2 + \sum_{t=1,t\neq f}^{T-1} w_t||\boldsymbol{y}_{t,m} - \boldsymbol{o}_{t,m}||_2^2\right) \tag{6}$$

Note that $\boldsymbol{y}_{f,m}$ and $\boldsymbol{o}_{f,m}$ are the label value and output of the facial point detection task, respectively. Additionally, the weight w_t is constant for each iteration. We update the CNN parameters $\boldsymbol{\theta}$, such as weight filter and connection weight, using backpropagation in Eq. (3).

Table 1. Parameters of heterogeneous learning CNN structure

Input	Image size	100 × 100
Layer 1	Filter size	9 × 9 × 16
	Maxout	2
	Max pooling	2×2
Layer 2	Filter size	9 × 9 × 32
	Maxout	2
	Max pooling	2×2
Layer 3	Filter size	9 × 9 × 64
	Maxout	2
	Max pooling	2×2
Layer 4	Sigmoid	200(Dropout:50%)
Output		17

4 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, we use regression estimation to detect five facial points : the left eye, right eye, nose, left mouth, and right mouth, 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 66. Race recognition is identified as Asian, White, or Black.

We employ a CNN that consists of three convolutional layers and three fully connectied layers, as shown in Tab. 1. In training of single task CNN, convolution layers and fully connected layers are have the same structure. In contrast, number of units in the output layer is equal to the number of classes for each facial attribute task. 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 the training sample consists of 42,663 images and the test sample consists of 11,000 images. The input images are 100×100 grayscale. We will publish this facial attribute dataset as soon as it is ready. The evaluation method of facial point detection is the same as that of Dantone et 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%.

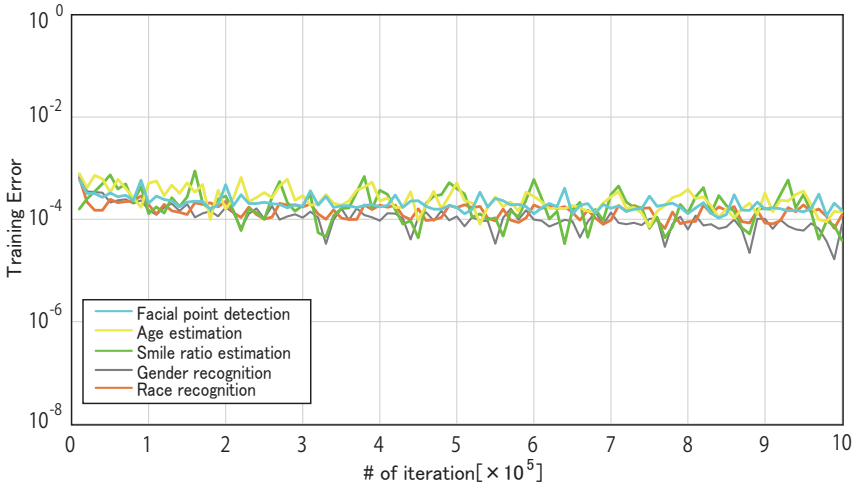

Fig. 4. Comparison of the training error for proposed method

4.1 Comparison of training errors

Figure 4 shows the training error for each task for the proposed method. The training error for conventional heterogeneous learning is different from the convergence value 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.

4.2 Comparison of performance for the proposed method

In Fig. 5, we compare the performances of single task learning conventional heterogeneous learning and the proposed method. The accuracy of the regression tasks is lower for single task learning than conventional heterogeneous learning, especially for the facial point detection task. This is because facial point detection most negatively affects other tasks training errors, such as the difference between training error variation. Compared with conventional heterogeneous learning and the proposed method, we improved performance by approximately 5% and the accuracy of the facial point detection task by approximately 14%. This means that the proposed method is stably able to train by extracting available facial features.

Figure 6 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. Additionally, the left images are the input image and result of facial point detection by the conventional

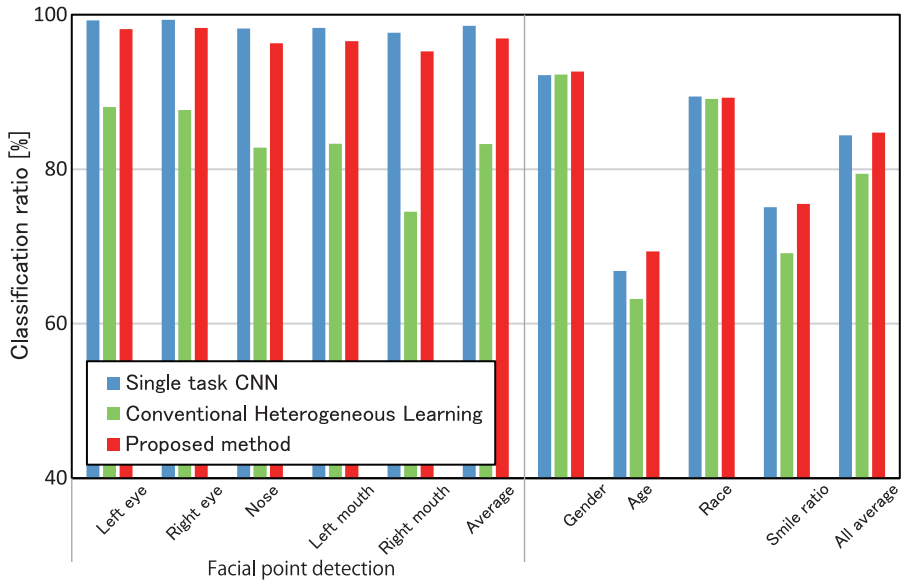

Fig. 5. Comparison of the proposed method and other method

heterogeneous learning or 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. 6, 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.

5 Disscusion

In this section, we define the effectiveness of the proposed method by comparing various viewpoints of conventional heterogeneous learning and the proposed method.

5.1 Performance of integrating training error functions

In conventional heterogeneous learning, mean square error function and cross entropy error function are employed as error functions of the regression and recognition tasks, respectively. However, there are noticeable differences between the error ranges of the mean squared error function and the cross entropy error

Fig. 6. Comparison of examples of facial image analysis

function. 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. In this section, we evaluate integrating the error range by changing training error functions.

Figure 7 shows the experimental results of classification accuracy that integrates mean square error function or not. When we compare the accuracy of regression tasks, proposed method is improved performance by approximately 20% . This mean that we can suppress the difference of error between regression tasks and recognition tasks, and this way can improve the accuracy of regression tasks that are susceptible to affect the training error of recognition tasks, especially.

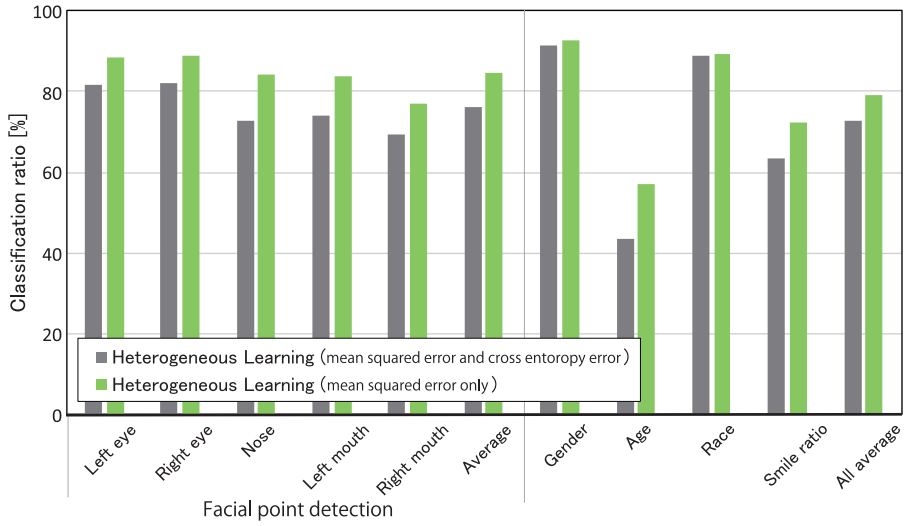

Fig. 7. Results of accuracy that integrates mean square error function or not

5.2 Regression tasks performance when threshold shifts

In experiments at section 4.2, we evaluate regression tasks that set to be fixed threshold, and if the output of a regression task is over than the threshold, the output is correct. if the output of regression task is under than the threshold, the output is missing classification. Therefore, we evaluate the accuracy of regression tasks by shifting the threshold for each method.

Figure 8 shows classification accuracy that shifts the threshold between 5 to 20 for facial point detection. If we compare the CNN of single task, proposed method is less performance than CNN of single task. However, If we compare the heterogeneous learning, proposed method is significantly better performance

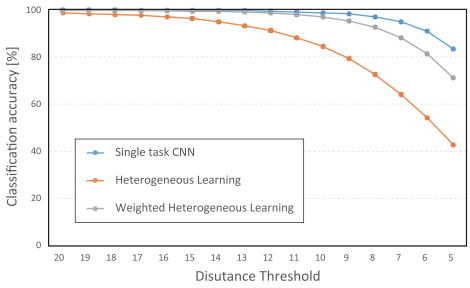

Fig. 8. Classification accuracy of facial point detection

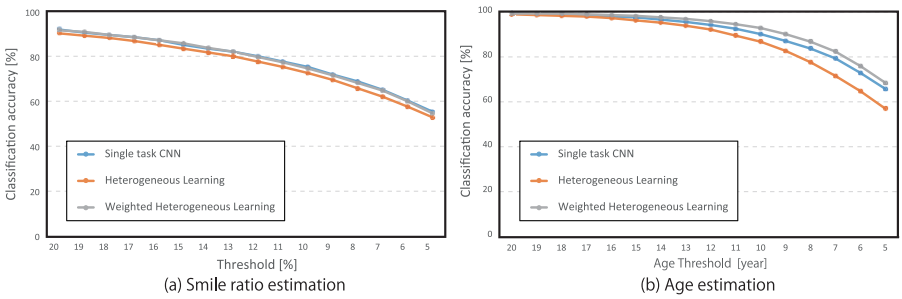

Fig. 9. Classification accuracy of smile ratio and age estimations

than conventional heterogeneous learning. Figure 9(a) and Fig. 9(b) show classification accuracy that shifts the threshold between 5 to 20 for smile ratio estimation and age estimation, respectively. Proposed method is better performance than conventional heterogeneous learning and CNN of single task in age estimation. Improving the performance of age estimations was caused by improving the facial point detection indebted proposed method. When CNN is trained with facial position by using heterogeneous learning, CNN easily focuses on facial part, and improving the performance by getting features that effectual estimate.

5.3 Visualization weight filters and feature maps

Weight filters and feature maps of CNN with heterogeneous learning are visualized in Figure 10. Note that, we visualize them in the first layer. Figure 10(a) shows visualization of weight filters and feature maps of conventional heterogeneous learning, and Fig. 10(b) shows visualization of weight filters and feature maps of the proposed method. Weight filters of conventional heterogeneous learning are shown a clear contrast between light and shape, as shown in Fig. 10(a). However, conventional heterogeneous learning was outputted weak response at facial part such as eye and mouth in feature maps. On the other hand, weight

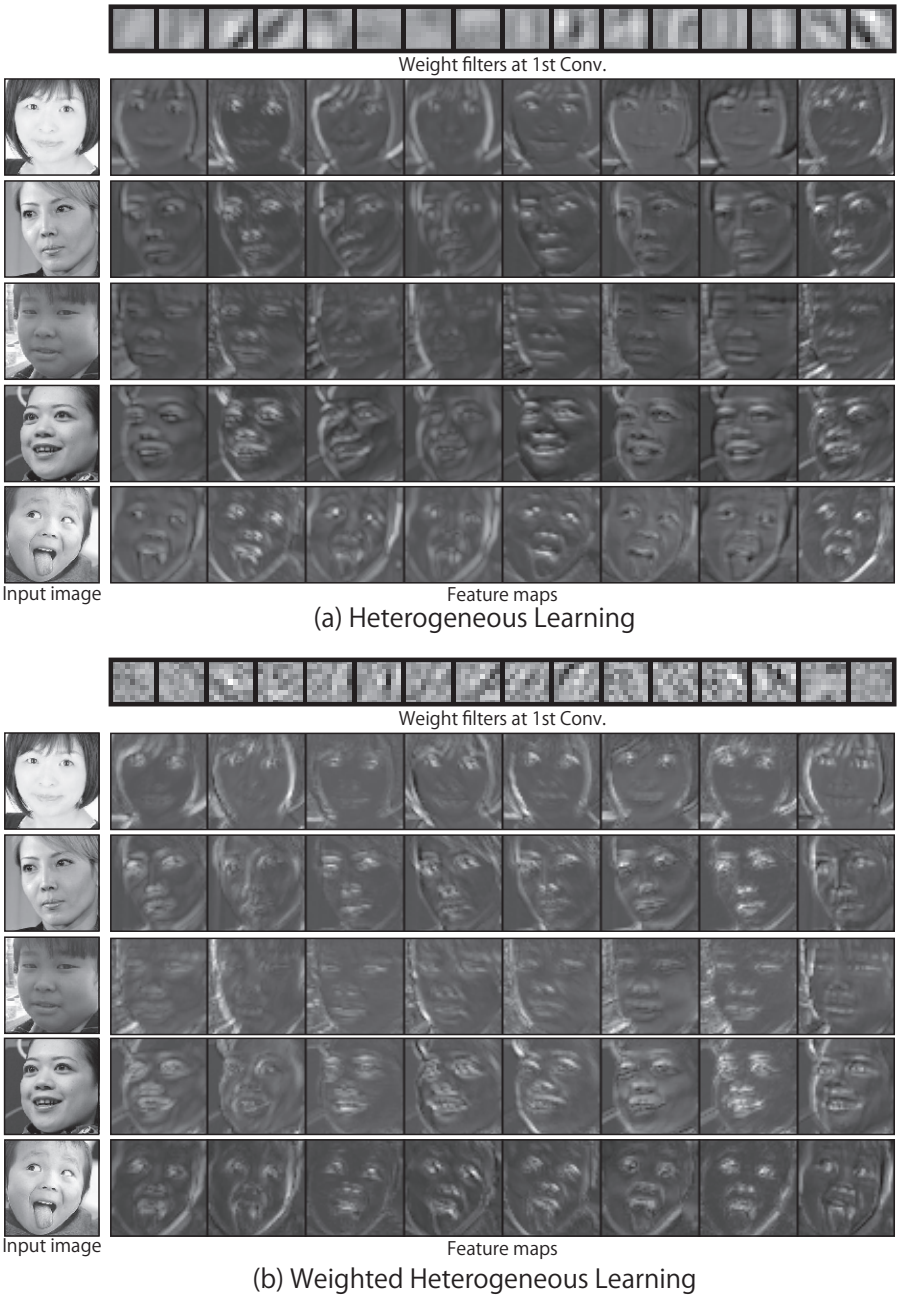

Fig. 10. Visualization weight filters and feature maps

filters of proposed method were noisy, nevertheless, proposed method was outputted strong response at facial part such as eye and mouth in feature maps, as shown in Fig. 10(b).

6 Conclusion

In this paper, we proposed a method to improve the performance of heterogeneous learning for facial image analysis. As a result, compared with conventional heterogeneous learning, 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. In European Conference on Computer Vision, 1998.
2. M. Dantone, J. Gall, G. Fanelli, and L. V. Gool. Real-time facial feature detection using conditional regression forests. In Computer Vision and Pattern Recognition, 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. H. C. Lian, and B. L. Lu. Multi-view Gender Classification Using Local Binary Patterns and Support Vector Machines. Advances in Neural Networks, Vol. 3972, pp. 202-209, 2006.
6. A. Krizhevsky, S. Ilva, G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Network. In Advances in Neural Information Processing System 25, pp.1097-1105, 2012.
7. Y. Sun, X. Wang, and X. Tang. Deep Convolutional Network Cascade for Facial Point Detection. In Computer Vision and Pattern Recognition, 2013.
8. Z. Erjin, F. Haoqiang, C. Zhimin, J. Yuning, Y. Qi. Extensive Facial Landmark Localization with Coarse-to-Fine Convolutional Network Cascade. IEEE International Conference on Computer Vision Workshops, 2013.
9. T. Yamashita, T. Watasue, Y. Yamauchi, and H. Fujiyoshi. Facial Point Detection Using Convolutional Neural Network Transferred from a Heterogeneous Task. International Conference on Image Processing, 2015.
10. M. Kimura, T. Yamashita, Y. Yamauchi, and H. Fujiyoshi. Facial point detection based on a convolutional neural network with optimal mini-batch procedure. IEEE International Conference on Image Processing, 2015.
11. W. Yue, and J. Qiang. Discriminative Deep Face Shape Model for Facial Point Detection. International Journal of Computer Vision, 2015.
12. A. Jourabloo, and X. Liu. Large-pose Face Alignment via CNN-based Dense 3D Model Fitting. Computer Vision and Pattern Recognition, 2016.

13. C. Yan, C. Lang, T. Wang, X. Du, and C. Zhang. Age Estimation Based on Convolutional Neural Network. In Pacific-Rim Conference on Advances in Multimedia Information Processing, Vol. 8879, pp. 211-220, 2014.
14. E. Eidinger, R. Enbar, and T. Hassner. Age and Gender Estimation of Unfiltered Faces. In IEEE Press Transactions on Information Forensics and Security, 2014.
15. Y. Zhu, L. Yan, M. Guowang, and G. Guodong. A Study on Apparent Age Estimation. IEEE International Conference on Computer Vision Workshops, 2015.
16. K.Zhanghui, H. Chen, Z. Wei. Deeply Learned Rich Coding for Cross-Dataset Facial Age Estimation.IEEE International Conference on Computer Vision Workshops, pp. 96-101, 2015.
17. F. H. C. Tivive , and A. Bouzerdoum. A Gender Recognition System using Shunting Inhibitory Convolutional Neural Networks. In Neural Networks, pp. 5336-5341, 2006.
18. 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.
19. G. Levi, and T. Hassner. Age and Gender Classification Using Convolutional Neural Networks. Computer Vision and Pattern Recognition, 2015.
20. A. Andreas, E. Theodoros, and P. Massimiliano, "Convex Multi-task Feature Learning", Kluwer Academic Publishers, Vol. 73, No. 3, pp. 243-272, 2008.
21. Z. Zhang, P. Luo, C. C. Loy, X. Tang. Facial Landmark Detection by Deep Multi-task Learning. in Proceedings of European Conference on Computer Vision, 2014.
22. D. Terrance, B. Kumar, W. T. Graham. Multi-task Learning of Facial Landmarks and Expression. Computer and Robot Vision, 2014.
23. Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Learning Deep Representation for Face Alignment with Auxiliary Attributes. arXiv preprint arXiv:1408.3967, 2015.
24. Y. Junho, J. Heechul, Y. ByungIn, C. Chngkyu, P. Dusik, and K. Junmo. Rotating Your Face Using Multi-task Deep Neural Network. Computer Vision and Pattern Recognition, 2015.
25. R. Ranjan, M. P. Vishalx, and R. Chellappa. Facial Landmark Detection by Deep Multi-task Learning. arXiv preprint arXiv:1603.01249, 2016.
26. D. E. Rumelhart, G. E. Hinton, R. J. Williams. Learning representations by back-propagating errors. In Neurocomputing, pp. 696-699, 1988.