17th International Conference on Machine Vision Applications (MVA)
Fully Online, July 25–27, 2021.

Facial landmark detection transfer learning for a specific user in driver status monitoring systems

Jaechul Kim†, Kensuke Taguchi†, Yusuke Hayashi†, Jungo Miyazaki†, Hironobu Fujiyoshi‡
†Advanced Technology Research, Institue Minatomirai Research Center
Kyocera Corporation, Yokohama Japan, ‡Chubu University, Aichi Japan
{jaechul.kim.yb, kensuke.taguchi.xm, yuusuke.hayashi.zs, jungo.miyazaki.zs}
@kyocera.jp†, fujiyoshi@isc.chubu.ac.jp‡

Abstract

The wide variety of human faces make it nearly impossible to prepare a complete training data set for facial landmark detection. Because of this, the performance of facial landmark detection is unlikely to be sufficient for driver status monitoring (DSM) systems. To improve the performance for a specific person (SP) by collecting data about that person, we propose the generator and discriminator model using the Lucas–Kanade assistance (GDA) algorithm for compiling a training data set. Even when data for a specific user can be collected, another issue is how to efficiently, effectively, and quickly re-train the model using an insufficient data set. To address this problem, we propose a novel method of transfer learning in the context of composite backbone networks (CBNet). The assistant backbone of CBNet is trained on a large unspecified people (USP) data set in the source domain and transfers its representation to the lead backbone, which is trained by a small SP data set in the target domain. In addition, we design an assistance loss function with output that is not only close to the SP data set, but also consistent with a USP data set with respect to labeled images. We test the proposed method using the 300 Videos in the Wild (300VW) data set and our own data set. Furthermore, show that the proposed method improves the stability of predictions. We expect our method to contribute to the realization of stable DSM systems.

1 Introduction

Facial landmark detection is used for various applications in driver status monitoring (DSM) systems, such as face recognition[15], gaze estimation[10], face tracking[8], and facial expression recognition[9]. In each case, application performance strongly depends on the results of facial landmark detection. In particular, if there are no data similar to a specific user in the training data set, the generated heatmaps will be inaccurate and infer incorrect detection points. Conversely, if data similar to a specific user exist in the learning data, a more accurate heat map is generated, improving the results (Figure 1). We consider that the two most important problems in facial landmark detection algorithms for DSM systems are how to collect user-specific training data that improve performance, and how to re-train a model using a collected specific person (SP) data set efficiently. In this paper, we propose a generator and discriminator model using the Lucas-Kanade (LK) assistance (GDA) algorithm for generating SP data set labels automatically with no iteration algorithm. Current high-performance algorithms[18, 16] improve performance by repeating algorithms or through parallelization using multiple networks. Better performance is desirable but increases the time and resources required for learning. For efficient model re-training, we design a CBNet[20] transfer learning approach, which is a combination of CBNet and transfer learning. These algorithms are very similar conceptually. CBNet is an assistance backbone that assists the lead backbone to obtain better features for the classifier. In transfer learning, a source domain dataset helps a target domain to learn efficiently. The USP data set is the assistance backbone for CBNet and the source domain for transfer learning. The SP data set is the lead backbone for CBNet and the target domain for transfer learning. Our main contributions are as follows:

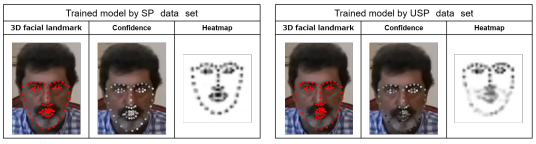

Figure 1. The confidence of the detection results in images is expressed in black and white density. Brighter indicates higher confidence, and darker indicates lower confidence.

1. We propose the GDA algorithm, which generates an SP data set automatically in a DSM system. In order to obtain better labels, the GDA algorithm proposes an ensemble of semi-supervised and unsupervised labels.

2. Our method can re-train a model using an insufficient data set and decreases the time required to train a model by CBNet transfer learning. We design an assistance loss function with output that is not only close to the SP data set, but also consis-

tent with the USP data set with respect to labeled images.

3. We evaluate our method using the 300VW[5, 2, 14] data set. It is confirmed that our method is more stable than existing methods.

2 Related work

This work is closely related to homogeneous transfer learning and semi-supervised learning, in that we apply the GDA algorithm to obtain a user-specific training data set, and we apply CBNet transfer learning to efficiently train a user-specific model.

2.1 Facial landmark semi-supervised learning

Several methods for pseudo labeling of unlabeled data in semi-supervised scenarios have been proposed[19, 7, 3]. Semi-supervised learning and unsupervised learning methods have been proposed for facial landmark detection[18, 16, 11, 8]. We propose the GDA algorithm, which reduces the number of iterations to one and the number of networks to one. To compensate for the performance degradation due to this simplification, we add an LK tracking algorithm.

2.2 Homogeneous transfer learning

In homogeneous transfer learning, the source and target domain have the same task and feature space. This method focuses on how to bridge gaps in data distributions between the source and target domains. This method has not yet been applied to facial landmark detection. In the proposed method, CBNet transfer learning is used for model training. A model trained with a USP data set acts as the source domain, because the data set contains supervised data and the model has learned about many people using a much larger amount of data. In contrast, the target domain learner is a model trained with an SP data set.

3 Methodology

To compile the SP data set, we first acquire video data from which we can automatically generate label information by the GDA algorithm. We train our model using CBNet transfer learning to reduce learning time and utilize the USP data set. Furthermore, we designed assistance loss for improving performance.

3.1 Label generation by GDA algorithm

The GDA algorithm plays the three roles of generator, discriminator, and assistance (Figure 2). The purpose of this algorithm is to create good quality SP labels. The generator generates pseudo labels in the videos data set, the discriminator judges the quality of pseudo labels from the generator, and the assistance which consists of the LK forward and backward tracking algorithm makes the final decision regarding pseudo label quality. Only labels passing that decision are used as pseudo labels in the SP data set.Generator and discriminator network are based on Dong's work[18], and assistance approach is inspired by Dong's work[16].I would appreciate Dong's previous work.

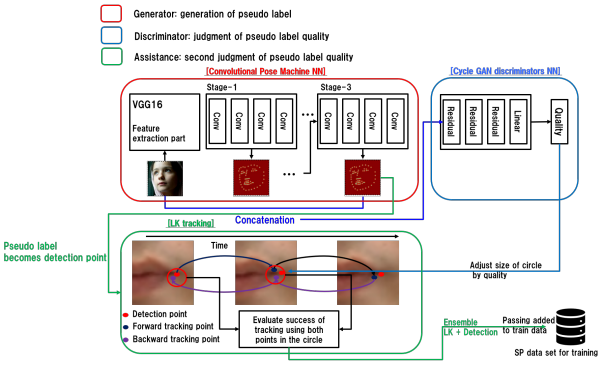

Figure 2. Video or data about only one person in the 300VW data set become the input of the discriminator and assistance. The result of discriminator is quality of label which is generated by generator. By number of quality from discriminator, adjust size of circle which assistance use for evaluation success.

3.2 CBNet transfer learning

When only a model pretrained by the USP data set is used, the performance improvement is expected to be limited. During re-training of the model using the SP data set as new data, the model becomes fine-tuned to the SP data set. Consequently, the model loses a lot of general USP data set information. To avoid this limitation, we propose CBNet transfer learning, which is a combination algorithm of CBNet and transfer learning. The assistant backbone of CBNet is trained by a large USP data set in the source domain and transfers its representation to a lead backbone, which is trained by a small SP data set in the target domain. This method has considerable potential for improving performance though the lead backbone network becoming more powerful for extracting features for a classifier (Figure 3).

3.3 Assistance loss

Assistance loss can be calculated by taking the differential between SP_detection (Figure 3) and USP_detection (Figure 3). A model trained by the USP data set has better general performance because it contains fully supervised labels and much more data than

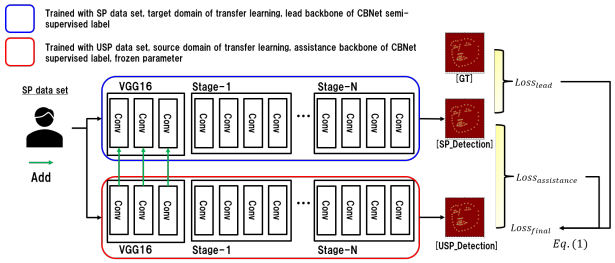

Figure 3. The assistance backbone assists the lead backbone when training the model with the SP data set. For using $loss_{final}$ by Equation 1 allowing better performance.

the SP data set. However, a model trained by the SP data set probably lacks accurate labels because it is not generated in a supervised manner. Assistance loss helps the model achieve learning performance similar to that of the SP data set while including labeled images as in the USP data set (Figure 3). We therefore consider that assistance loss mitigates inaccuracies in the SP data set, thereby realizing more stable model training and improved performance of facial landmark detection. $loss_{final}$ is calculated as

$$loss_{final} = \sum_{n=1}^{N} loss_{lead}^{n} + \gamma \sum_{n=1}^{N} loss_{assistance}^{n} \qquad (1)$$

where $loss_{lead}$ is calculated by the mean squared error, $loss_{assistance}$ is the L2 norm, and N is the number of facial landmarks. We tested several values for the weighting factor γ, and found the best performance with a value of 0.2.

4 Experiment

In this section, we present the results of our experiments on the 300VW data set. We evaluate detection performance using the Normalized Mean Error (NME)[6, 12, 13, 17], and the Area Under the Curve (AUC) [1, 4].

4.1 Experimental data setup

The evaluation SP data set contains data with supervised labels that do not pass the GDA algorithm, and the training SP data set contains data with labels assigned by the generator that pass the GDA algorithm (Figure 4).

4.2 Result

Proposed method, the data collected by the GDA algorithm is a training data set based on inference. For GDA algorithm to compensate for inaccuracies,

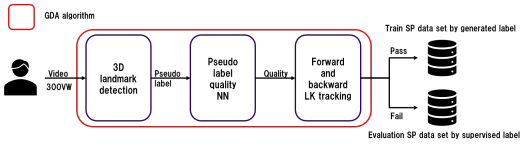

Figure 4. 300VW data without label is applied to the GDA algorithm. Pass data added to SP train data. And fail data was used as evaluation data.

tracking points obtained by the forward–backward LK tracking algorithm and the inferred results were combined (Figure 2). And model is trained by SP data set using CBnet transfer learning with assistance loss. Our propose method gave more stable performance than conventional method(Figure 5).

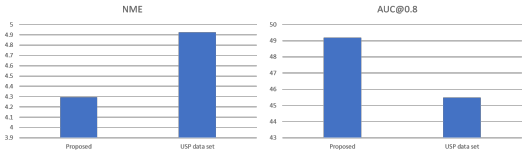

Figure 5. Our proposed method achieves better performance than does the conventional method.

4.3 Qualitative Comparisons

Whereas the predictions from models trained by a USP data set are often failures, the proposed method achieves more stable results (Figure 6).

Figure 6. Qualitative results in 300VW. The first two rows show predictions for lips on a bearded face, whereas the results from the proposed method give more stable predictions.

4.4 Automatically generated data set

Figure 7 shows the results from a model trained using the proposed method with our own automatically generated data. The performance is better than when using a model trained using the USP data set.

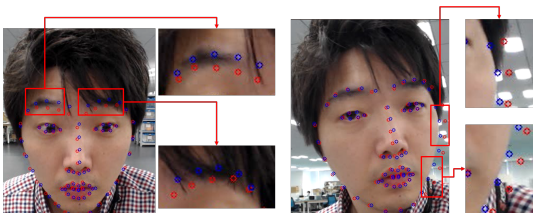

Figure 7. Blue circles indicate detection by a model trained with our own data using the proposed method and red circles indicate detection by a model trained with a USP data set. The proposed method can produce more stable predictions.

4.5 Comparison of learning time

Learning using the proposed method converged after 20 epochs, whereas existing methods converged after 50 epochs (Figure 8). The difference comes from the different amount of training data between the existing method and proposed method. The existing method was trained using 60K items of USP data which is whole 300VW data except for SP data, while the proposed method was trained using under 2K items of SP data generated by the GDA algorithm. As a result, learning using the proposed method converged 33 times faster.

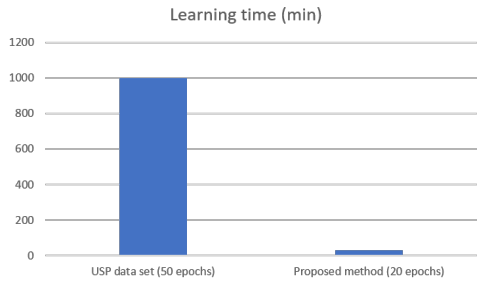

Figure 8. Comparison of learning time between the conventional and proposed methods. The proposed method converges much faster.

5 Ablation study

Through experiments, we confirmed the results of learning using the small amounts of SP data set collected by the GDA algorithm was better than learning with large amounts of USP data set. By collecting SP data set efficient model learning became possible (Figure 9). In addition to training model using CBNet transfer learning performed better than using only an SP data set (Figure 9). This method can train model quickly. And our propose method, after collecting SP data set by GDA algorithm, training model using CBNet transfer learning with assistance loss method achieved the best performance (Figure 9). Assistance loss can train model effectively by more suitable loss for model.

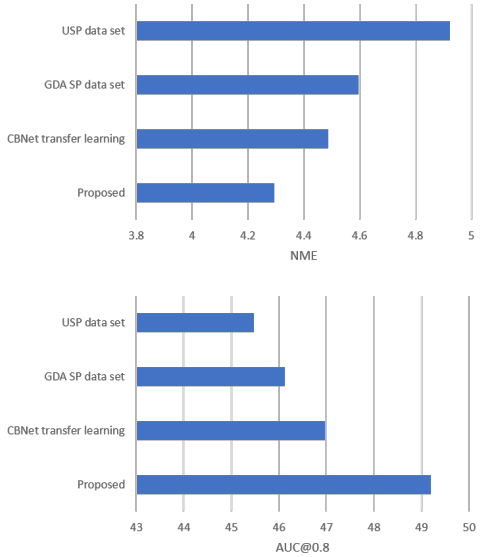

Figure 9. The results confirmed that learning with GDA SP data set gave better performance than USP data set. Furthermore, by training model using CBNet transfer learning performance improved. And proposed method, train model using CBNet transfer learning with assistance loss was even more improved.

6 Conclusion

The proposed method automatically compiles a training data set using the GDA algorithm for an SP in a limited resource environment, and we confirmed that the model can be trained efficiently using the CBNet transfer learning method with assistance loss. The time required to train the model and the cost of compiling the training data set were far lower than with conventional methods, and performance was improved.

References

[1] Bulat A. and Tzimiropoulos G. How far are we from solving the 2d & 3d face alignment problem?(and a

dataset of 230,000 3d facial landmarks). *ICCV*, 2017.

[2] Sagonas C., Tzimiropoulos G., Zafeiriou S., and Pantic M. 300 faces in-the-wild challenge: The first facial landmark localization challenge. *ICCVW*, 2013.

[3] Ma F., Meng D., Xie Q., Li Z., and Dong X. Self-paced co-training. *ICML*, 2017.

[4] Trigeorgis G., Snape P., Nicolaou. M. A., Antonakos. E., and Zafeiriou S. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. *CVPR*, 2016.

[5] S. Zafeiriou G. S. Chrysos, E. Antonakos and P. Snape. Offline deformable face tracking in arbitrary videos. *ICCVW*, 2015.

[6] Lv J., Shao X., Xing J., Cheng C., and Zhou X. A deep regression architecture with two-stage reinitialization for high performance facial landmark detection. *CVPR*, 2017.

[7] Pawan M., Benjamin K., and Koller P. Daphne. Self-paced learning for latent variable models. *NeurIPS*, 2010.

[8] Khan M. H., J. McDonagh, and Tzimiropoulos G. Synergy between face alignment and tracking via discriminative global consensus optimization. *ICCV*, 2017.

[9] Munasinghe M. I. N. P. Facial expression recognition using facial landmarks and random forest classifier. *IEEE/ACIS*, 2018.

[10] Park S., Zhang X., Bulling A., and Hilliges O. Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. *ACM*, 2018.

[11] Qian S., Sun K., Wu W., Qian C., and Jia J. Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation. *ICCV*, 2019.

[12] Ren S., Cao. X, Wei. Y, and Sun. J. Face alignment via regressing local binary features. *IEEE TIP*, 2016.

[13] Zhu1 S., Li C., Loy C., and Tang X. Unconstrained face alignment via cascaded compositional learning. *CVPR*, 2016.

[14] G. Tzimiropoulos. Project-out cascaded regression with an application to face alignment. *CVPR*, 2015.

[15] Liu W., Wen Y., Yu Z., M. Li, Raj B., and Song L. Sphereface: Deep hypersphere embedding for face recognition. *ICCV*, 2017.

[16] Dong X., Yu S., Weng X., Wei S., Yang Y., and Sheikh Y. Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. *CVPR*, 2018.

[17] Dong X., Yan Y., Ouyang W., and Yang Y. Style aggregated network for facial landmark detection. *CVPR*, 2018.

[18] Dong X. and Yang Y. Teacher supervises students how to learn from partially labeled images for facial landmark detection. *ICCV*, 2019.

[19] Bengio Y., Louradour J., Collobert R., and Weston J. Curriculum learning. *ICML*, 2009.

[20] Liu Y., Wang Y., Wang S., Liang T., Zhao Q., Tang Z., and Ling H. Cbnet: A novel composite backbone network architecture for object detection. *arXiv:1909.03625*, 2019.